SNP-set Tests for Sequencing and Genome-Wide Association Studies

SNP-set Tests for Sequencing and Genome-Wide Association Studies

A dissertation presented

by

Ian James Barnett

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2014

Dissertation Advisor: Professor Xihong Lin Ian James Barnett

SNP-set Tests for Sequencing and Genome-Wide Association Studies

Abstract

In this dissertation we propose methodology for testing SNP-sets for genetic associations, both for sequencing and genome-wide association studies. Due to the large scale of this kind of data, there is an emphasis on producing methodology that is not only accurate and powerful, but also computationally efficient.

In the Chapter 1, we aim at using extreme phenotype sampling to increase the power to identify rare variants associated with complex traits. We confirm both analytically and numerically that sampling individuals with extreme phenotypes can enrich the presence of causal rare variants and can therefore lead to an increase in power compared to random sampling. While application of traditional rare variant association tests to these extreme phenotype samples requires dichotomizing the continuous phenotypes before analysis, the dichotomization procedure can decrease the power by reducing the information in the phenotypes. To avoid this, we propose a novel statistical method based on optimal SKAT (SKAT-O) that allows us to test for rare variant effects using continuous phenotypes in the analysis of extreme phenotype samples. The increase in power of this method is demonstrated through simulation of a wide range of scenarios as well as in the triglyceride data of the Dallas Heart Study.

In Chapter 2, we present the *higher criticism*, a signal detection method that is effective for testing the joint null hypothesis against a sparse alternative, in the context of SNP-set testing. This test is useful for testing the effect of a gene or a genetic pathway that consists of d genetic markers. Accurate p-value calculations for the higher criticism based on the asymptotic distribution require a very large d, which is not the case for the number of genetic variants in a gene or a pathway. We propose an analytic method that

accurately computes the p-value of the higher criticism test for finite d problems. Unlike previous treatments of the higher criticism, this method does not rely on asymptotics in d or simulation, and is exact for arbitrary d when test statistics are normally distributed. The method is also particularly computationally advantageous when d is not large. We illustrate the proposed method with a case-control genome-wide association study of lung cancer and compare its power to competing methods through simulations.

In Chapter 3, we adapt the higher criticism to better allow for correlation in the SNP-set. In Chapter 2, the SNPs in the SNP-set are first decorrelated, which loses power. We propose the generalized higher criticism (GHC) that does not require asymptotics in the number of SNPs in the SNP-set while simultaneously allowing for arbitrary correlation structures among the SNPs in the SNP-set. The detection boundary of the test is obtained, and the power of this method is compared with existing SNP-set tests over simulated regions with varied correlation structures and signal sparsity. The relative performance of these methods is also compared in their analysis of the CGEM breast cancer genome-wide association study.

Contents

Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies

Ian Barnett, Seunggeun Lee and Xihong Lin

Department of Biostatistics

Harvard School of Public Health

1.1 Introduction

With the increase in the number of sequencing studies (Biesecker et al., 2011), there is a newfound access to samples with low minor allele frequency (MAF $1-5\%$) and rare (MAF $< 1\%$) genetic variants. In the search for genetic components of complex traits, discovered common variants (MAF $> 5\%$) from genome-wide association studies explain only a small proportion of the total heritability of these traits (Ioannidis et al., 2009; Maher, 2008; Manolio et al., 2009). As a result, attention has turned to low frequency and rare variants instead expecting that they could play an important role in uncovering gene-phenotype relationships (Cirulli and Goldstein, 2010; Ji et al., 2008; Nejentsev et al., 2009; Ng et al., 2008; Ramser et al., 2008). Unfortunately, rare variants are difficult to detect in even reasonably large samples. This problem can be alleviated through the development of powerful study designs. To this effort, numerous association studies have chosen to sample subjects with extreme phenotypes in the hope of increasing power to detect causal SNPs (Clément et al., 1995; Gu et al., 1997; Hu et al., 2009; Khor and Goh, 2010; Li and Leal, 2008; Liang et al., 2000; Price et al., 2010; Risch and Zhang, 1995). There have also been numerous developments in methodology to detect QTLs under these extreme phenotype sampling (EPS) study designs(Chen et al., 2005; Huang and Lin, 2007; Li et al., 2011; Slatkin, 1999; Wallace et al., 2006). A fundamental assumption that motivates these EPS methods is that rare causal variants are more likely found in the extremes of the quantitative trait. In this paper, we support the use of this practice by showing both analytically and numerically that EPS increases the presence of rare causal variants in a variety of settings. As a result, we show that EPS is more powerful for detecting for rare variant effects than random sampling.

Various methods have been proposed to tackle the challenge of association testing for rare variants. Burden tests such as the Combined Multivariate and Collapsing method (CMC) (Li and Leal, 2008), Cohort Allelic Sums Test (CAST) (Morgenthaler and Thilly, 2007) and the Weighted Sum Test (WST) (Madsen and Browning, 2009) combine information from all rare variants within a target region such as an exon or gene by collapsing them into a single genetic variable, which is tested for association with the phenotypes of

interest. Numerous rare variants testing methods have been developed using the same strategies (Bansal et al., 2010; Basu and Pan, 2011; Lee et al., 2012a; Morris and Zeggini, 2010; Price et al., 2008). A limitation of all burden tests is that they could lose significant amount of power in the presence of variants with different association directions and a large fraction of non-causal variants in the region. Alternatively the Sequence Kernel Association Test (SKAT) (Wu et al., 2011a) aggregates evidence of individual variant effects across the region using a kernel function and uses a computationally efficient mixed model variance component test to test for association. SKAT can naturally adjust covariates and has robust power in the presence of variants with different association directions and a large proportion of null variants. It is also a generalization of several non burden tests such as C-alpha test (Neale et al., 2011), the SSU test (Pan, 2009), and the haplotype association test (Tzeng and Zhang, 2007). Recently the optimal SKAT (SKAT-O) (Lee et al., 2012b) has been proposed to unify the burden test and SKAT to a single framework and to construct the optimal test within the framework.

Moreover, limited statistical methods have been developed for studying rare variant effects when extreme phenotypes are sampled. In a typical EPS study, the two extremes are treated as two different groups representing a dichotomous phenotype. For example, Hu et al. (Hu et al., 2009) used the contrast between subjects with high HDL-C levels against those with low HDL-C levels to identify an association with the ABCA1 gene. If the same method of extreme sampling were to instead retain the continuous phenotype values, the gain in information could provide greater power to detect gene-phenotype associations. For common variants, Huang and Lin (Huang and Lin, 2007) proposed testing for associations between extreme continuous phenotypes and variants using the maximum likelihood method assuming a truncated normal distribution for extreme phenotype. Recently, this approach was adapted by Li et al. (Li et al., 2011) to accommodate testing for multiple rare variant effects with the burden CMC approach. As a burden test, this approach is powerful when most variants in a region are causal and the effects of causal variants are in the same direction. However, it loses power in the presence of variants with different association directions or a large number of non-causal variants in a region.

In this paper, we first confirm both analytically and empirically that EPS substantially increases the chance to observe rare causal variants and hence increases their observed frequencies in finite study samples. Using this result, we demonstrate that EPS provides a more powerful design strategy for testing rare variant effects compared to random sampling. We next develop a new more powerful statistical method for testing for rare variant effects in EPS. Specifically, we extend SKAT and the optimal SKAT (SKAT-O) to EPS by analyzing extreme phenotypes as continuous variables within a likelihood framework. We show that the proposed tests perform well in a wide range of situations and outperform burden tests. We further show that analysis using continuous extreme phenotypes (CEP) improves power for detecting rare variant effects compared to using dichotomized extreme phenotypes (DEP). We illustrate the finite sample performance of proposed methods by conducting extensive simulations and application to analysis of triglyceride levels from the Dallas Heart Study (Victor et al., 2004).

1.2 Material and Methods

1.2.1 Goals and notation

The goal is to find an optimum sampling strategy when resources are limited and to develop powerful association test methods to detect phenotype-genotype associations. We evaluate the effectiveness of extreme phenotype sampling (EPS) compared to random phenotype sampling.

We first confirm analytically that extreme phenotype sampling enriches causal rare variants by increasing their MAFs (Appendix A.1). We consider the cases with a single causal variant and multiple causal variants and calculate the MAF in extreme phenotype sampling as a function of the population MAF, the threshold used to select extreme phenotypes, and the effect sizes of genotypes.

We next evaluate the two different methods that utilize EPS phenotypes in different ways: the method that retains continuous phenotypes and the method that dichotomizes them into cases and controls. We consider the case with a sample of n individuals who have been sequenced in a genomic region of interest containing p genetic variants. The

ith individual has covariate information over m covariates $\boldsymbol{X}_i = [X_{i1}, , X_{im}]^T$, genotypes of the p variants in the region $\boldsymbol{G}_i = [G_{i1}, , G_{ip}]^T$, and a continuous phenotype y_i. The genotype G_{ij} represents the number of copies of the minor allele of the jth variant that the ith individual has.

1.2.2 Model

To test for an association between the variants and continuous phenotype while controlling for covariates, consider a linear model

$$y_i = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{X}_i + \boldsymbol{\beta}' \boldsymbol{G}_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. Here α_0 is an intercept term, $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_m]'$ is a vector of regression coefficients for p genetic variants. The null hypothesis of $H_0 : \boldsymbol{\beta} = 0$ corresponds to no genetic effect on the trait. Since a p-DF likelihood ratio test has little power to detect causal variants particularly in the presence of a large number of rare variants, the gene-phenotype relationship is instead tested for by region-based tests such as burden tests and non-burden tests, e.g., SKAT. An adaptation of the CMC (Li and Leal, 2008) burden test is used that collapses genotype information by counting the number of variants in the region before applying logistic regression to the collapsed statistic. We call this test DEP-Burden.

1.2.3 Association tests under the extreme phenotype sampling design

Since both SKAT and burden tests are capable of handling dichotomous phenotypes in the case-control setting, they can be applied to test for associations after using EPS. Dichotomizing the high phenotypic extremes as cases and the lower phenotypic extremes as controls is a natural extension of each tests functionality. However, applying SKAT and SKAT-O to continuous phenotype data obtained from EPS requires further development, since the extreme continuous phenotypes do not follow Gaussian distribution due to the phenotypic selection. Suppose we select n samples with either $y_i > c_1$ or $y_i < c_2$, and denote the selected y_i as y_i^*. Then under the null hypothesis y_i^* follows

truncated Gaussian distribution with a density function

$$f(y_i^*) = \frac{\phi(\boldsymbol{X}_i\alpha, \sigma^2)}{\Phi(c_2, \sigma^2) + 1 - \Phi(c_1, \sigma^2)}$$

where $\phi(\mu, \sigma^2)$ and $\Phi(\mu, \sigma^2)$ are density and distribution functions of the Gaussian distribution with mean 0 and variance σ^2.

To increase test power and decrease the test DF, we assume β_j follows an arbitrary distribution with mean 0 and variance ψw_j^2. We note that $H_0 : \boldsymbol{\beta} = 0$ is equivalent to $H_0 : \psi = 0$. The score test statistic of $\psi = 0$ is

$$Q_s = \sum_{j=1}^{p} w_j^2 \left(\sum_{i=1}^{n} G_{ij}(y_i^* - \hat{\mu}_i) \right)^2$$

where $\hat{\mu}_j$ is an estimated mean of y_i^* under the null hypothesis. We now show that Q_s asymptotically follows a mixture of chi-squares distribution.

1.2.4 The null distribution of Continuous Extreme Phenotype SKAT

Suppose the null model is true where $\boldsymbol{\beta} = 0$ and where $\boldsymbol{X}_i = [x_{i0}, ..., x_{im}]'$ are covariates of the ith individual with $x_{i0} = 1$ and $\boldsymbol{\alpha} = [\alpha_0, ..., \alpha_m]'$ is the vector of regression coefficients of $\boldsymbol{X}_i$ and $\epsilon_i \sim N(0, \sigma^2)$. Suppose we select n samples with either $y_i > c_1$ or $y_i < c_2$, and denote the selected y_i as y_i^*. Under the null hypothesis, y_i^* follows a truncated Gaussian distribution with density function

$$f(y_i^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{-(y_i^* - \boldsymbol{X}_i'\boldsymbol{\alpha})^2/2\sigma^2}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})}$$

where $t_{i1} = (c_1 - \boldsymbol{X}_i'\boldsymbol{\alpha})/\sigma$ and $t_{i2} = (c_2 - \boldsymbol{X}_i'\boldsymbol{\alpha})/\sigma$. The first derivative of the log-likelihood function is

$$u_j = \frac{dL}{d\alpha_j} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{X}_i'\boldsymbol{\alpha} + m_i)$$

and the second derivative is

$$J_{jk} = \frac{d^2L}{d\alpha_j d\alpha_k} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}x_{ik}(-1 + v_j)$$

where

$$m_i = \sigma \frac{\phi(t_{i2}) - \phi(t_{i1})}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})}, \text{ and } v_i = \frac{t_{i2}\phi(t_{i2}) - t_{i1}\phi(t_{i1})}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})} + \frac{m_i^2}{\sigma^2}$$

Define $\boldsymbol{S} = -\boldsymbol{J}$, $\boldsymbol{y}^* = [y_1^*, ..., y_n^*]'$, $\boldsymbol{u} = [u_0, ..., u_m]'$, and $\boldsymbol{m} = [m_1, ..., m_n]'$. By Fisher Scoring (or Newton Raphson) procedure, a new $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha}^* = \boldsymbol{\alpha} + \boldsymbol{S}^{-1}\boldsymbol{u}$$

hence $\boldsymbol{S}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) = \boldsymbol{u}$. Since $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}/\sigma^2$, where $\boldsymbol{V} = diag\{(1 - v_i)\}$ and $\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_n]'$,

$$\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) = \boldsymbol{X}^*(y^* - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{m})$$

Now treat the Fisher Scoring algorithm as a weighted least squares problem. Define the working vector

$$\tilde{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{V}^{-1}(y^* - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{m})$$

,

and then $\boldsymbol{\alpha}^*$ is a weighted least squares estimator of the linear model $\tilde{y} = \boldsymbol{X}\boldsymbol{\alpha} + \tilde{\epsilon}$ with $E(\tilde{\epsilon}) = 0$ and $Var(\tilde{\epsilon}) = \boldsymbol{V}^{-1}$. Since $E(y_i^*) = \boldsymbol{X}_i^*\boldsymbol{\alpha} - m_i$, the SKAT test statistic with linear weighted kernel is

$$\begin{aligned} Q_s &= (y^* - \hat{\mu})'\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'(y^* - \hat{\mu}) \\ &= (\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\alpha}^*)'\boldsymbol{V}\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'\boldsymbol{V}(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\alpha}^*) \\ &= \tilde{\boldsymbol{Y}}\boldsymbol{P}_0\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'\boldsymbol{P}_0\tilde{\boldsymbol{Y}} \end{aligned}$$

where $\boldsymbol{P}_0 = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}$. From $Var(\tilde{y}_i) = (1 - v_i)\sigma^2$, the asymptotic null distribution of Q_s is

$$\sum \lambda_v \chi_v^2$$

where λ_v is the vth eigenvalue of $\hat{\sigma}^2 \boldsymbol{P}_0^{1/2} \boldsymbol{G} \boldsymbol{W} \boldsymbol{G}' \boldsymbol{P}_0^{1/2}$.

Calculations of Q_s require fitting the null model using extreme phenotypes y_i^* under truncated normal likelihood. The Newton-Raphson method can be used to estimate α and σ^2. P-values can be obtained by approximating the mixture of chi-square distributions with a non-central chi-square distribution by matching he moments or inverting the characteristic function (Davies, 1980).

1.2.5 Optimal Continuous Extreme Phenotype SKAT

Recently Lee et al. (2012) (Lee et al., 2012b) proposed an optimal unified approach, which unifies SKAT and burden test to adaptively select best test structure, called SKAT-O. We can not only extend SKAT, but also SKAT-O to continuous extreme phenotypes. Suppose QB is the score test statistics of the weighted burden test:

$$Q_B = \left(\sum_{i=1}^{n} (y_i^* - \hat{\mu}_i) \sum_{j=1}^{p} w_j G_{ij} \right)^2$$

and then the test statistic of unified test is

$$Q_\rho = (1-\rho) Q_s + \rho Q_B$$

where ρ ($0 \leq \rho \leq 1$) is a parameter to determine whether test is close to SKAT ($\rho = 0$) or burden tests ($\rho = 1$). It is based on a recent generalization of SKAT which allows the correlation among variants effects βs. Under this setting, they proposed the optimal SKAT, called SKAT-O. This test is defined by selecting the ρ that minimizes the p-value of the SKAT-O test statistic,

$$T = \min_{0 \leq \rho \leq 1} p_\rho$$

where p_ρ is a p-value with given ρ. The test statistic T can be obtained by simple grid search across a range of ρ: set a grid $0 = \rho_1 < \rho_2 < \cdots < \rho_b = 1$, then $T = \min(p_{\rho_1}, ..., p_{\rho_b})$. In simulation studies and real data analysis, we used the equal sized grid of 11 points (from 0 to 1) to obtain T. From the fact that the Q_ρ can be decomposed to the shared random variables, asymptotic p-value of T can be obtained through

computationally efficient one-dimensional numerical integration (Appendix A.2). We use this extreme phenotype optimal SKAT in our simulation studies and data analysis.

When the sample size is small, SKAT family methods (including SKAT and SKAT-O) can produce conservative results with both binary and extreme continuous phenotypes. To resolve this issue, Lee et al. (Lee et al., 2012b,a) have proposed a method to adjust asymptotic null distribution by estimating small sample moments when the trait is dichotomous. We employ a similar approach (Appendix A.3). For all simulation studies and real data analysis we used small sample adjustment for SKAT methods given the small to moderate sample sizes we considered. We used SKAT-O for continuous extreme phenotype SKAT (CEP-SKAT-O) and dichotomous extreme phenotype SKAT (DEP-SKAT-O), and for random sample continuous phenotypes (RS-SKAT-O). It should be noted that for larger sample sizes, the small sample adjustment is not necessary. Through simulations we found sample sizes lower than n=500 to benefit from the small sample adjustment, with sample sizes as low as n=1000 to not benefit from the adjustment.

1.2.6 Type I error simulations

We first generated haplotype data by the forward simulator, SFS_CODE (Hernandez, 2008), which offers the ability to incorporate purifying selection on deleterious variants and thus provides better model to simulate variants in exomes. Data were simulated according to the European demographic model with a population bottleneck followed by exponential growth. We simulated 32,000 haplotypes each 100,000 base pairs wide as our population base. To achieve a simulated sample over a 3kb exon, a random 3kb region is selected (containing 41 variants on average) and each individual genotype is formed by combining at random two haplotypes over that region. Phenotypes for the i-th individual in a sample were produced from the generated genotype and covariate data according to

$$Y_i = 0.5X_{i1} + 0.5X_{i2} + \epsilon_i$$

Where the covariate X_{i1} is 1 with probability 0.5 and 0 otherwise, and the covariate X_{i2} and the residual ϵ_i are both instances of a standard normal random variable.

Using the simulated genotype and phenotype data for the N individuals, a random sample of size n is selected. For random sampling of continuous traits, SKAT-O with the default $w_j = Beta(1, 25)$ weight is used to test for an association between the continuous phenotype and genotype after controlling for both covariates, producing a p-value (RS-SKAT-O) . In order to test for the association between variants and phenotype under EPS using the standard dichotomizing method, we treat the highest ($n/2$) extremes as cases and lowest ($n/2$) extremes as controls. The dichotomized phenotypes are used by both DEP-SKAT-O and DEP-Burden. This same extreme phenotype sample is used to compare with the tests that retain the continuous phenotype (CEP-SKAT-O and CEP-Burden). A p-value for the CEP-SKAT-O test and the CEP-Burden burden test are produced from these continuous phenotype values and the corresponding genotype and covariate data. The proportion of p-values below a specified α-level provides an estimate for the Type I error at that significance level.

1.2.7 Power simulations

Power comparisons between the various sampling methods were performed using simulated genotype data as was used in the Type I error simulation setting. After generating the genotypes for N individuals, 20% of the variants with $MAF < 0.03$ are selected to be causal variants. Different percentages of causal variants were also considered. Phenotypes are then generated for the N individuals according to:

$$y = 0.5X_1 + 0.5X_2 + \beta_1 G_1 + ... + \beta_p G_p + \epsilon$$

The covariate X_1 is generated as a Bernoulli random variable with $p = 0.5$. The covariate X_2 and the added noise ϵ are generated independently from a standard normal distribution. Non-causal variants are assigned $\beta_j = 0$, and the causal variants are generated according to:

$$|\beta_j| = -a\log_{10}(MAF_j)$$

Here, $a > 0$ is a parameter that specifies the strength of variant-phenotype associations, hence the strength of heritability. Large values of a lead to stronger effects of causal variants on phenotype and cause rare variants to become more enriched in the phenotypic

extremes. In one simulation setting an increase in a from 0.3 to 0.4 increases the heritability of the phenotype from 0.034 to 0.042. The heritability also increases with the number of causal variants. To obtain an estimate of the heritability, the proportion of the variance in phenotype explained by the genotypes of causal variants is estimated assuming no LD between variants.

Power estimates are obtained for various (extreme phenotype) sample sizes (n=500, 1000, and 2000), percentages in each phenotypic extreme sampled (10% and 25%), percentages of causal variants with a positive effect (80%, 100%), and percentages of causal variants with $MAF < 0.03$ (20%, 40%, and 60%).

1.3 Results

1.3.1 Extreme sampling enriches rare causal variants

Our analytical calculations (See Material and Methods and Appendix A.1) confirm that rare causal variants can be enriched in phenotypic extremes. The degree of enrichment increases when more extreme phenotypes are sampled and a higher percentage of causal variants are present in a region. To empirically validate this finding, randomly selected 3kb exonic regions were simulated using the population genetic simulation model with European demographic history (see Material and Methods). For each 3kb region, causal variants were randomly selected to be 100%, 70%, 40% and 0% of sufficiently rare variants ($MAF < 0.03$) and the jth causal variant was given the effect size β_j as a function of its MAF. Note that these causal variant percentages differ from those in the power simulations so as to further accentuate the effect of causal variant percentage on the inflation of MAF due to EPS. Also for the power simulations, causal variant percentages of 10% and 20% were used instead. Phenotypes are then generated from a linear model with heritability of genetic variants being 2.6%, 1.3%, and 0%.

Because the causal variants are known in the simulation setting, the expected MAF of a causal variant using EPS can be computed analytically (Appendix A.1). The expected MAFs of causal variants using EPS matched closely with the sample MAFs of causal variants using EPS (Figure 1.1). The MAFs of simulated causal variants after EPS had an

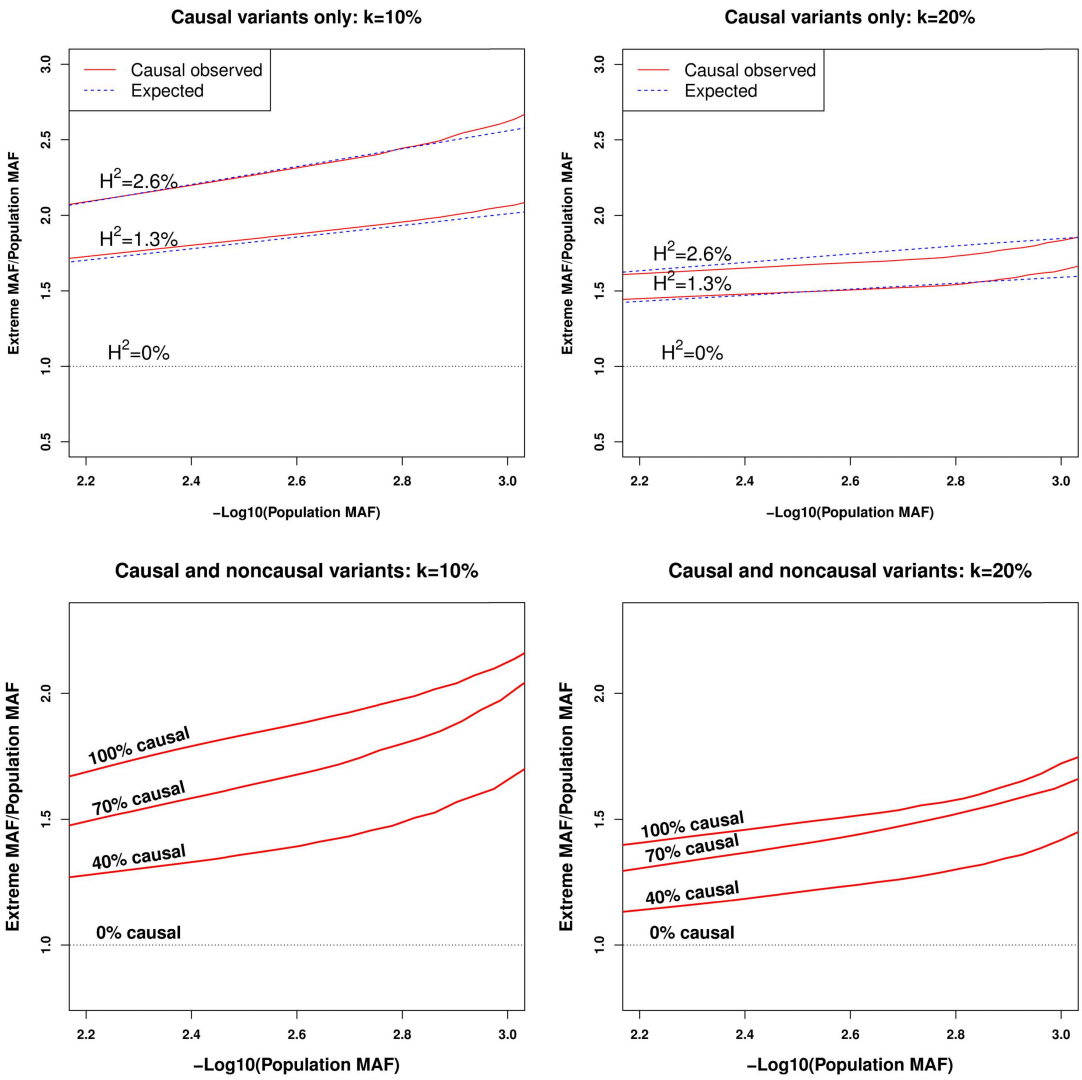

Figure 1.1: Estimated folds increase of the observed MAFs of causal variants in phenotypic extremes over population MAFs. The red lines represent the smoothed observed fold increases. The dotted lines represent the theoretical fold increase. For each causal variant, population MAF was computed using the full simulated population while extreme phenotype MAF was computed after sampling the tails. See Appendix A.1 for derivation of theoretical expected MAF for extreme phenotypes. The top two figures consider the case where all variants are causal by sampling $k = 10\%$ and 20% high/low extremes. For each case, three situations were considered by heritability of causal variants: $H^2 = 2.6\%, 1.3\%$, and 0% (no causal variant). Higher heritability gives more enrichment of rare variants. The bottom two figures consider the case where different fractions of variants in a region are causal (100%, 70%, 40% and 0%) by sampling $k = 10\%$ and 20% high/low extremes. Presence of non-causal variants in a region lower the degree of enrichment of rare variants.

overall increased frequency over the respective population MAFs. This trend decreases as samples are restricted to less extreme phenotypes and heritability is lower. No enrichment is found when there is no causal variant. When both the causal and non-causal variants in a region are considered simultaneously, the median MAF using EPS is much less inflated than when only causal variants are examined.

1.3.2 Sampling methods for comparison

Motivated by the enrichment of causal rare variants in phenotypic extremes, we expect to find that EPS methods can increase power to detect rare causal variants over random sampling methods. We extend the SKAT family methods to test for region-level rare variant effects when continuous phenotypes obtained from EPS are used in analysis. In simulation and data analysis, we only use the extreme phenotype optimal SKAT (SKAT-O), which accounts for extreme phenotype sampling and unifies the burden test and SKAT to a single framework and by constructing the optimal test within the framework. Using the simulated genotype data over 3kb regions the phenotypes were generated using the additive linear model (see Material and Methods). Given the same sample size, we compare the power of three tests designed for detecting rare variant effects using EPS. We first consider a burden test, DEP-Burden, that uses dichotomized extreme phenotypes along with collapsed information over genotypes by simply counting the number of rare variants with $MAF < 3\%$ in the gene before applying logistic regression to the collapsed statistic. We also apply this same collapsed statistic to continuous extreme phenotypes as done in Li et al. (Li et al., 2011) and call this test CEP-Burden. Next we consider dichotomized extreme phenotype SKAT-O (DEP-SKAT-O), which applies optimal SKAT (SKAT-O) to dichotomized extreme phenotypes while applying small sample adjustments when sample sizes are small (Lee et al., 2012a). Finally we consider continuous extreme phenotype SKAT (CEP-SKAT-O), which does not dichotomize and instead extends linear regression optimal SKAT over the continuous extreme phenotypes (see Material and Methods) by using a truncated normal distribution. We also applied the small sample adjustment to CEP-SKAT-O to obtain the correct type I error rates when sample sizes are small. To demonstrate the benefits of EPS compared to random sampling, we included in

the comparison a fourth method using random sampling SKAT-O (RS-SKAT-O), which applies optimal SKAT to the continuous phenotypes of a random sample. We assume the same sample size when comparing different methods so their powers are comparable. The power of each competing method is estimated as the proportion of p-values less than $\alpha = 10^{-6}$ in an effort to imitate genome-wide association studies.

The type 1 error rates for CEP-SKAT-O were accurate at $\alpha = 0.01$ and $\alpha = 0.05$ and slightly inflated at genome-wide significance levels $\alpha = 10^{-6}$(see Table A.1). When all causal variants had the same direction of effect, CEP-SKAT-O and CEP-Burden had the greatest power with a substantial lead over every other method (Figure 1.2). When causal variants had effects in opposite directions all tests lost power uniformly due to less enrichment of rare variants, but CEP-SKAT-O became the most powerful by a large margin (Figure 1.3). In this case DEP-Burden had the least power. The power of the three methods employing SKAT-O (CEP-SKAT-O, DEP-SKAT-O, and RS-SKAT-O) is much more robust to changes in the proportion of causal variants that have a positive effect than the burden tests power is. This is because SKAT-O allows for each individual variant to affect phenotype in different directions and also allows for no effect. On the other hand, burden tests assume all the causal variants share the same direction of effect and that all the variants in a region are causal, and so the power of the burden tests greatly diminishes when causal variants are allowed effects in opposite directions or many causal variants are allowed no effect.

When all causal variants having the same direction of effect and as the percent of rare variants that were causal increased, the power gap between DEP-Burden and DEP-SKAT-O reduces, an observation that is also true for the relationship between CEP-Burden and CEP-SKAT-O (Figure 1.2). This is because CEP-SKAT-O includes CEP-Burden as a special case and behaves like CEP-Burden automatically when most variants are causal with effects in the same direction. To see this, we found that in simulations the estimated ρ decreased by a factor of 0.36 on average when changing from the case of having all positive causal variant effects to the case of having causal variant effects in opposite directions.

Methods utilizing extreme sampling benefit as the cutoff for extreme phenotypes

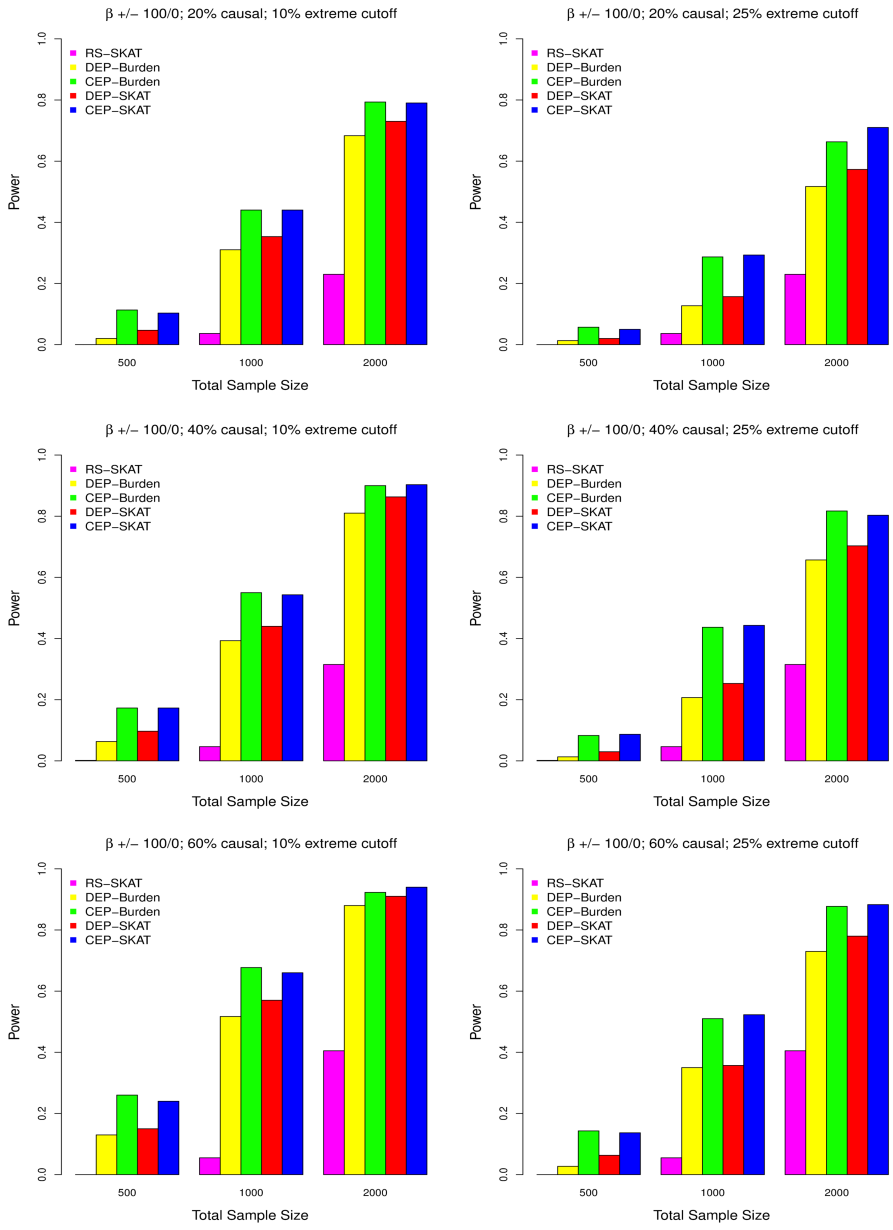

Figure 1.2: Simulated power comparisons between four rare variants association tests with all causal variants having a positive effect on phenotype. The five tests are random sample optimal SKAT (RS-SKAT-O), dichotomized extreme phenotype burden test (DEP-Burden), continuous extreme phenotype burden test (CEP-Burden), dichotomized extreme phenotype optimal SKAT (DEP-SKAT-O), and continuous extreme phenotype optimal SKAT (CEP-SKAT-O). The left panel considers the situation where 10% high/low extremes are sampled with the three rows corresponding to 20% (0.6% heritability), 40% (1.2% heritability) and 60% (1.8% heritability) variants in a 3kb region being causal. Three total sample sizes are considered: $n = 500$, 1000, 2000. The right panel considers the situation where 25% high/low extremes are sampled. Exonic regions are simulated with effect sizes for each causal variant equal to $\beta = -0.2\log_{10} MAF$. Power is estimated by the proportion of tests that detect an association at the $\alpha = 10^{-6}$ level.

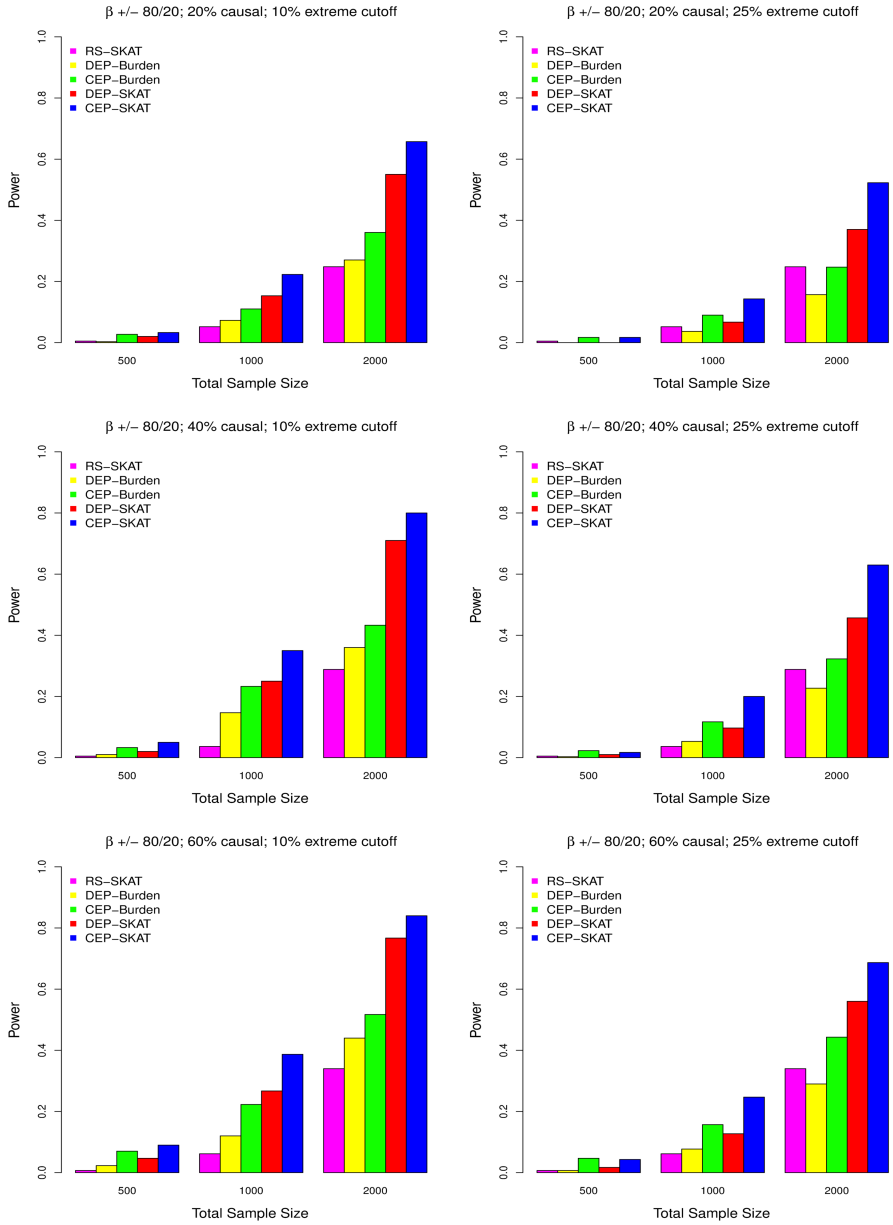

Figure 1.3: Simulated power comparisons between four rare variants association tests with 80% of rare causal variants selected to have a positive effect on phenotype while the remaining 20% have a negative effect. The five tests are random sample SKAT (RS-SKAT-O), dichotomized extreme phenotype burden test (DEP-Burden), continuous extreme phenotype burden test (CEP-Burden), dichotomized extreme phenotype optimal SKAT (DEP-SKAT-O), and continuous extreme phenotype optimal SKAT (CEP-SKAT-O). The left panel considers the situation where 10% high/low extremes are sampled with the three rows corresponding to 20% (0.6% heritability), 40% (1.2% heritability) and 60% (1.8% heritability) variants in a 3kb region being causal. Three total sample sizes are considered: $n = 500, 1000, 2000$. The right panel considers the situation where 25% high/low extremes are sampled. Exonic regions are simulated with effect sizes for each causal variant equal to $|\beta| = -0.2\log_{10} MAF$ with the effect being negated 20% of the time. Power is estimated by the proportion of tests that detect an association at the $\alpha = 10^{-6}$ level.

increases. In particular, as the percent of the tails sampled from the distribution of phenotype decreases from 25% to 10%, all EPS tests show incremental increases in power given the same sample size due to higher enrichment of rare variants. Relative power comparisons remain unchanged after decreasing the heritability of the phenotype and after increasing the exon length to 5kb or 10kb (regions of these lengths contain 69 variants and 138 variants on average, respectively). Also, simulations were also performed where the β was selected to be a constant rather than being a decreasing function of the MAF but the relative power of the methods remained the same (Figure A.3). Regardless of the setting, CEP-SKAT-O was consistently robust and had the greatest overall power to detect gene-phenotype associations over the other methods.

1.3.3 Application to the Dallas Heart Study data

In the Dallas Heart Study (Victor et al., 2004), 3476 individuals were sequenced over the genes ANGPTL3 (MIM 604774), ANGPTL4 (MIM 605910), and ANGPTL5 (MIM 607666). A total of 93 variants are present over these genes, and the variants in all three genes were tested simultaneously for an association with log-transformed serum triglyceride levels (logTG). Analysis for each of three genes separately is also considered (Appendix A.5). Ethnicity and sex were adjusted for in the analysis. To demonstrate rare variant association test methods for extreme phenotype sampling (EPS), a total of 1389 individuals with the highest 20% and lowest 20% of logTG levels in each age-gender stratum were selected as the EPS sample. The continuous values were used in CEP-SKAT-O while dichotomized values were used for DEP-SKAT-O and DEP-Burden. Random samples of equivalent size were selected for the RS-SKAT-O method for comparison purposes.

To compare the effects of the different cutoffs of tails, we considered to sample individuals from wider tails (30% and 40%). Since wider tails had more samples, to make p-values comparable, we randomly sub-sampled 1389 individuals among individuals in wider tails in order to have the same sample size as compared to a 20% cut off. In these cases, median p-values calculated from multiple random samples were obtained (Table A.2). The p-values for all EPS methods are sensitive to the extreme phenotype cutoff. CEP-SKAT-O outperforms the other methods when there is sufficient information about

the continuous trait distribution. It performs similarly to DEP-SKAT-O where there is limited information in the data, e.g., when the cutoff is low or when there are a small number of rare variants in a gene (Appendix A.5). When extremes were sampled from wider tails (30% and 40%) all of the tests tended to lose significance, demonstrating the strength of EPS. We also computed p-values with different cutoffs and unequal sample sizes (Figure A.2), and CEP-SKAT-O outperformed other competing methods overall.

1.3.4 Power estimation

In the planning of new sequencing studies, it is important to be able to estimate the power to detect causal variants under various study designs. We provide such power calculations for extreme phenotype sampling designs using CEP-SKAT-O. We use analytical formulas to obtain the distribution of our statistic by allowing users to specifying desirable parameters of interest (Appendix A.4). The parameters that can be specified by the user include sample size, the percent of causal variants, the length of the genomic region, the effect size of the causal variants, the proportion of causal variants that have a positive effect, and the proportion of the tails that are sampled in EPS. We find that power is increased as sample size increases, as the proportion of causal variants increases, as the effect size increases, when causal variants have their effect in the same direction, and when we are more selective by sampling individuals with more extreme phenotypes. The power is also dependent on the genomic region as the distribution of the number of genetic variants, the MAF distribution, and the LD structure vary over the genome, so for genome-wide studies power estimations are averaged over many randomly selected regions of equivalent size.

To evaluate the accuracy of these analytic power estimations, we show a side by side comparison with empirical power simulations (Figure 1.4 and Figure A.1). In this setting we consider 3kb regions with 20% of variants being causal with all effects in the same direction. We see that the estimated power with our analytic calculations matches the empirical power over a wide range of sample sizes.

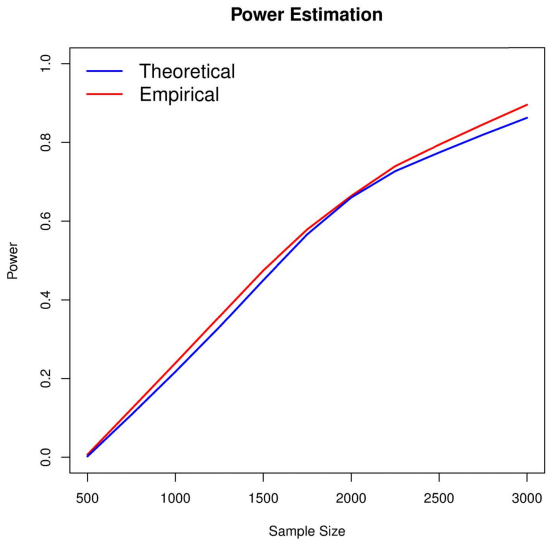

Figure 1.4: Estimated power of CEP-SKAT for testing 3kb regions with 20% of variants being causal with all effects in the same direction and the casual variants have effects to $|\beta| = -0.2\log_{10}MAF$. Theoretical power was calculated as described Appendix A.5, and empirical power was estimated by simulation using 300 replicates. No covariates were considered in either the theoretical or empirical power calculations. Furthermore empirical power was computed using CEP-SKAT without small sample adjustments.

1.4 Discussion

We confirm in this paper through analytical calculations and simulation studies that sampling phenotypic extremes of a population can enrich rare causal variants. As a consequence, we show that sampling from phenotypic extremes profit over analogous random sampling methods by showing a sizable gain in power when the same size is used. In particular, analysis using dichotomized extreme phenotype (DEP-SKAT-O) is shown to be more powerful than that using random sampling with continuous phenotypes (RS-SKAT-O) in almost all scenarios. We develop a new method, continuous extreme phenotype optimal SKAT (CEP-SKAT-O), which improves upon DEP-SKAT-O by retaining continuous phenotype information rather than dichotomizing, includes the continuous extreme phenotype burden test as a special case, and results in a significant increase in sensitivity to causal variants. We find that CEP-SKAT-O has the overall greatest power in a wide variety of settings over DEP-SKAT-O, RS-SKAT-O, and comparable

collapsing methods.

In the realm of region based association testing methodology, there already exist many methods capable of handling continuous phenotypes when a normal distribution is assumed. However in the case of extreme sampling, phenotype follows a truncated normal distribution instead, and so current methods cannot be directly applied without dichotomizing. The advantage of CEP-SKAT-O is that it adapts SKAT-O, a method that applies multiple linear regression of a phenotype on all genotypes in a region, to be able to handle phenotypes coming from a truncated normal. This adaptation allows for the usual continuous phenotype analysis without forcing the usual loss of phenotype information that occurs due to dichotomizing.

CEP-SKAT-O assumes that subjects are sampled from the extremes of a normally distributed phenotype, which in some circumstances may be inappropriate, and hence the test results can be biased when the normality of the underlying trait is violated. Dichotomizing phenotypes using DEP-SKAT-O is robust to departure from normality, although it is subject to some power loss when normality assumption is true. For candidate gene studies, permutation can be used to estimate the null distribution of the CEP-SKAT-O test statistic when the underlying trait does not follow a normal distribution. However, this is computationally difficult for GWAS where genome-wide significance levels are very stringent and a large number of genes are tested. We have found the maximum likelihood estimator of 2 seems to be sensitive to the distribution of the underlying trait. It is of future research interest to develop an alternative estimator of 2 that is more robust to deviations from normality.

In several ongoing exome sequencing studies conducted in the NHLBI Exome Sequencing Project, subjects were sampled using extremes of multiple phenotypes, future research is needed to develop methods for analyzing this more complex sampling setting. When subjects with extreme phenotypes are sampled for sequencing, covariate confounding effects need to be accounted for at the design phase to ensure representative samples. One strategy is to use stratified sampling, i.e. sample extreme phenotypes within key covariate stratum. For example, a phenotype distribution is likely to be gender-specific. It is desirable to sample extreme phenotypes within each gender stratum. The residual

covariate confounding effects can be adjusted for at the analysis stage.

Analytic P-value calculation for the higher criticism test in finite d problems

Ian Barnett and Xihong Lin

Department of Biostatistics

Harvard School of Public Health

2.1 Introduction

The higher criticism is a method to test for the joint null hypothesis against the alternative hypothesis that signals in a set are sparse. This situation is commonly encountered in genetic association studies, where it is often of interest to jointly test the effects of genetic variants within a gene, network, or pathway on a disease/trait (Tzeng and Zhang, 2007; Wu et al., 2011b). The higher criticism adaptively aggregates independent marginal test statistics, and has been shown to be an asymptotically powerful test of the joint null hypothesis when signals are sparse and are above the detection boundary (Donoho and Jin, 2004; Arias-Castro et al., 2011). Here "asymptotics" refers to the number of test statistics, d, tending towards infinity. The higher criticism test statistic is the supremum of a standardized empirical process under the null hypothesis and follows a Gumbel distribution asymptotically, but its convergence is very slow and its p-value cannot be reliably computed analytically based on asymptotic theory unless d is very large (Jaeschke, 1979).

However, many practical situations of interest that could benefit from the higher criticism do not have a very large d. For example, in genome-wide association studies, one is often interested in testing for the effect of genetic variants in a gene or a pathway. The number of genetic variants in a gene or a pathway is often not large, e.g., the number of genetic variants in a gene is often in the dozens for the vast majority of genes across the genome and the number of genetic variants in most genetic pathways is often in the hundreds. To test for the joint null hypothesis of no gene or no pathway effect, the asymptotic theory based p-values using the higher criticism are not applicable due to its very slow convergence rate. Simulation of the null distribution might seem a reasonable alternative to asymptotics. However, this is computationally burdensome in genome-wide association studies, as tens of thousands of genes of different sizes need to be tested, and a control for multiple comparisons results in very stringent significance levels. For example, in order to obtain p-values accurate to the genome-wide significance level of 10^{-7} for testing 10^4 genes requires at least a staggering 10^{11} total test statistics simulated under the null hypothesis.

We present in this paper an analytic method of accurate p-value calculations for the

higher criticism in non-large d signal detection settings. The proposed method relies neither on asymptotics in d nor on computationally expensive simulation of the null distribution. We show the proposed method is exact for an arbitrary d for normally distributed marginal test statistics, and is computationally fast for the non-large d settings commonly encountered in genome-wide association studies. We evaluate the finite sample performance of the proposed method using simulation and demonstrate the effectiveness of the method on data from a case-control lung cancer genome-wide association study.

2.2 The higher criticism and its asymptotic distribution

Consider d normally independent test statistics $Z = [Z_1, ..., Z_d]^T$ with means $\mu = [\mu_1, ..., \mu_d]^T$ and unit variance. It is of interest to test the joint null hypothesis that $H_0 : \mu = 0$ against the alternative that μ is a sparse vector with the number of non-zero entries $d_0 = d^{1-\beta}$ ($\beta \in (1/2, 1)$) (Donoho and Jin, 2004). Letting $\bar{\Phi}(t)$ be the survival function of the standard normal distribution and $S(t) = \sum_{j=1}^{d} I_{\{|Z_j| \geq t\}}$, the higher criticism test statistic is

$$HC = \sup_{t>0} \left(\frac{S(t) - d2\bar{\Phi}(t)}{[d2\bar{\Phi}(t)\{1 - 2\bar{\Phi}(t)\}]^{1/2}} \right). \tag{2.1}$$

Under the null, HC follows a Gumbel distribution as $d \to \infty$. For large d, gains in power can be made by searching for the supremum over a restricted range of t (Donoho and Jin, 2004). For $0 < \epsilon < \delta < 1$, if the supremum in (2.1) is taken over the interval $\Phi^{-1}(1 - \delta/2) < t < \Phi^{-1}(1 - \epsilon/2)$ then as shown in Jaeschke (1979), we can approximate the distribution of HC using its asymptotic distribution

$$\text{pr}(HC < c) \approx \exp[-\exp\{-c(2\text{log}\rho)^{1/2} - 2^{-1}\text{log}\pi + 2^{-1}\text{log}_2\rho + 2\text{log}\rho\}], \tag{2.2}$$

where $\rho = 2^{-1}\text{log}[\delta(1 - \epsilon)/\{\epsilon(1 - \delta)\}]$. Jaeschke (1979) shows that the higher criticism converges in distribution at an abysmal rate of $O\{(\text{log}d)^{-1/2}\}$. Our simulations show that the asmyptotic distribution is inaccurate for d as large as 10^6 (Appendix B.3). As a result, accurate higher criticism p-values at stringent significance levels for gene or pathway level analysis in genome-wide association studies need to be acquired without the use of

asymptotics.

In genetic association studies, the individual marker test statistics Z_j within a gene or a genetic pathway are often correlated with covariance Σ, which can be estimated from the genotypes of a study sample. Letting $UU^T = \Sigma$ be the Cholesky decomposition, then under the joint null hypothesis, the transformed statistics $U^{-1}Z$ are uncorrelated standard normal random variables and so the higher criticism can be applied directly on these transformed test statistics (Hall and Jin, 2010). This is appropriate only when sample size is larger than d, which is often the case in gene/pathway level analysis in genome-wide association studies.

2.3 Estimation of p-values for the higher criticism in finite d settings

The higher criticism test rejects the joint null hypothesis for large values of HC. We show in this section, finding the supremum does not require an exhaustive search over all $t > 0$. Let $HC(t) = \{S(t) - d2\bar{\Phi}(t)\}[d2\bar{\Phi}(t)\{1 - 2\bar{\Phi}(t)\}]^{-1/2}$. Then $HC(t)$ is a piecewise increasing function with a local maximum at each observed $|Z_i|$. Hence calculating the supremum in the higher criticism test statistic requires only finding a maximum over d quantities. Specifically, let h be the observed HC statistic in (2.1). Letting $c(t|h) = h[2d\bar{\Phi}(t)\{1 - 2\bar{\Phi}(t)\}]^{1/2} + 2d\bar{\Phi}(t)$, the p-value corresponding to a given observed higher criticism statistic h is

$$\text{pr}\left[\sup_{t>0}\{HC(t)\} \geq h\right] = 1 - \text{pr}\left[\bigcap_{t>0}\{S(t) < c(t|h)\}\right]. \tag{2.3}$$

Upon first glance, evaluating the p-value in (2.3) seems to require determining the probability of an intersection of an uncountable number of events (one for each $t > 0$). Without having asymptotics in d, we can instead leverage the fact that $S(t)$ is binomially distributed and can only take on a finite number of values $0, ..., d$. This will reduce the intersection in (2.3) to an intersection over a finite number of events as defined by the partition given in Lemma 1.

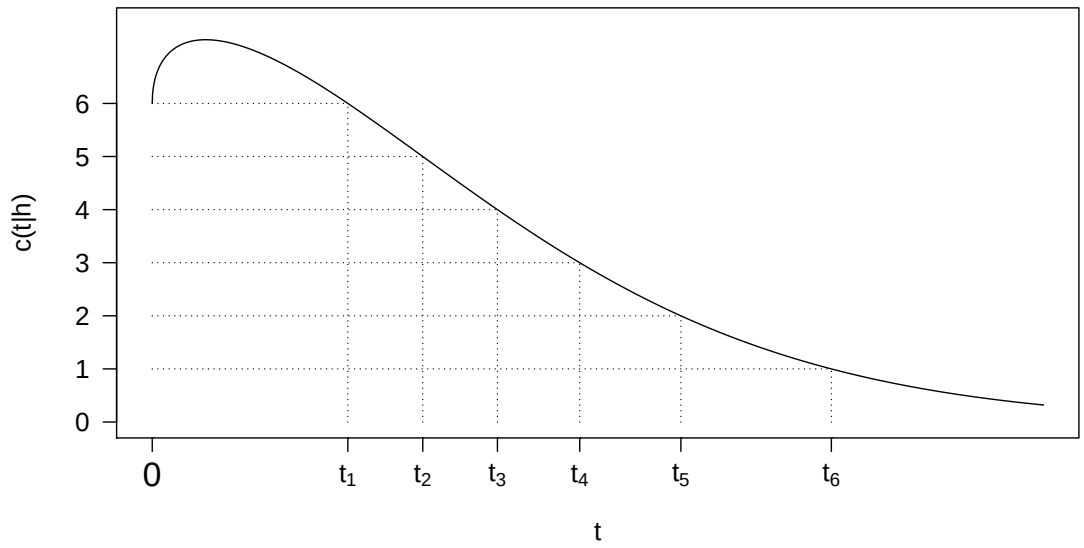

Figure 2.1: An example $c(t|h) = h[2d\bar{\Phi}(t)\{1 - 2\bar{\Phi}(t)\}]^{1/2} + 2d\bar{\Phi}(t)$ is plotted with $d = 6$ and $h = 2.4$. The partition given by Lemma 1 is labeled on the t-axis.

Lemma 1. *There exists a partition of the positive real line* $0 = t_0 < \cdots < t_{d+1} = \infty$ *such that* $c(t|h) > d$ *for* $t_0 < t < t_1$ *and* $d - k < c(t|h) \leq d - k + 1$ *for* $t_k \leq t < t_{k+1}$ *for each* $k = 1, ..., d$.

The proof of Lemma 1 is left to Appendix B.2. Lemma 1 makes the observation that $c(t|h)$ as a function of t takes a value of d when $t = 0$, and then increases to a global maximum before decreasing with an asymptote at 0 (Fig. 2.1). The form of $c(t|h)$ is the same for all $h > 0$ and so in each case the partition given by Lemma 1 exists.

We can ignore the case where the observed higher criticism statistic $h = 0$, because in this case the p-value trivially takes the value of 1. For $h > 0$, using Lemma 1, Theorem 1 simplifies the p-value expression in (2.3) to the joint probability of a finite set which is computationally feasible.

Theorem 1. *Letting* $0 = t_0 < \cdots < t_{d+1} = \infty$ *be the partition given by Lemma 1, then*

$$\text{pr}\left[\bigcap_{t>0}\{S(t) < c(t|h)\}\right] = \text{pr}\left[\bigcap_{k=1}^{d}\{S(t_k) \leq d - k\}\right].$$

The proof of Theorem 1 is left to Appendix B.1. According to Theorem 1 for the partition given in Lemma 1, the p-value expression in (2.3) simplifies to

$$1 - \text{pr}\left[\bigcap_{k=1}^{d}\{S(t_k) \leq d - k\}\right]. \tag{2.4}$$

For a given d and h, this partition is obtained by solving for t_k in the equation $c(t_k|h) = d - k + 1$ for each $k = 1, ..., d$. The result is

$$t_k = \Phi^{-1}\left[1 - \frac{2(d-k+1) + h^2 - h\{h^2 + 4(d-k+1) - 4(d-k+1)^2/d\}^{1/2}}{4(h^2+d)}\right], \qquad (2.5)$$

which defines the partition.

Evaluating (2.4) directly is difficult because each of the d events in the intersection are not independent. Instead, by the chain rule for conditioning the p-value can be written as

$$1 - \mathrm{pr}\left[\bigcap_{k=1}^{d}\{S(t_k) \le d-k\}\right] = 1 - \prod_{k=1}^{d}\mathrm{pr}\left[S(t_k) \le d-k \mid \bigcap_{l=1}^{k-1}\{S(t_l) \le d-l\}\right]. \qquad (2.6)$$

It is well known that empirical processes have the Markov property (Gaenssler, 1983). It follows that conditional on $S(t_{k-1}) = m_{k-1}, ..., S(t_1) = m_1$, then $S(t_k)$ is Binomial with m_{k-1} trials and probability of success $\bar{\Phi}(t_k)/\bar{\Phi}(t_{k-1})$; information about $S(t_{k-2})$,...,$S(t_1)$ has no bearing on the distribution of $S(t_k)$ if $S(t_{k-1})$ is known. We can utilize this to compute the terms in the product of (2.6) by further conditioning on $S(t_{k-1})$ in the kth term. Letting $q_{k,a} = \mathrm{pr}\{S(t_k) = a \mid S(t_{k-1}) \le d-k+1, ..., S(t_1) \le d-1\}$, some calculations show that

$$\begin{aligned} q_{k,a} &= \sum_{m=0}^{d-k+1}\mathrm{pr}\{S(t_k) = a \mid S(t_{k-1}) = m\}\frac{q_{k-1,m}}{\sum_{l=0}^{d-k+1} q_{k-1,l}} \\ &= \sum_{m=0}^{d-k+1} I_{\{a\le m\}}\binom{m}{a}\{\bar{\Phi}(t_k)/\bar{\Phi}(t_{k-1})\}^a\{1-\bar{\Phi}(t_k)/\bar{\Phi}(t_{k-1})\}^{m-a}\frac{q_{k-1,m}}{\sum_{l=0}^{d-k+1} q_{k-1,l}}. \end{aligned} \qquad (2.7)$$

From (2.7), in order to compute $q_{k,a}$, only knowledge of $q_{k-1,m}$ for $m = 0, ..., d-k+1$ is required. Because $q_{1,a} = \mathrm{pr}\{S(t_1) = a\}$ is a binomial probability, this offers a base case for calculating the p-value by computing each $q_{k,a}$ for $k = 1, ..., d$ and $a = 0, ..., d-k$.

The main result of this paper, Theorem 2, integrates the $q_{k,a}$s into the exact analytic p-value calculation of the higher criticism for an arbitrary d.

Theorem 2. *For the partition in* (2.5)*, the* $q_{k,a}$*s from* (2.7)*, and the observed higher criticism*

	d		
α	2	10	50
5·0e-2	4·98e-2(5·95e-2)	4·98e-2(3·24e-2)	5·01e-2(1·84e-2)
1·0e-2	1·00e-2(2·24e-2)	9·92e-3(7·31e-3)	1·01e-2(1·59e-3)
1·0e-3	9·97e-4(2·05e-3)	1·01e-3(6·03e-4)	9·75e-4(4·90e-5)
1·0e-4	1·01e-4(5·32e-4)	1·12e-4(7·30e-5)	9·80e-5(4·00e-6)

Table 2.1: Estimated type I error rates from 10^6 simulations for the higher criticism using the proposed analytic method over a range of significance levels, α. Asymptotic type I error rates are provided for comparison: exact(asymptotic).

statistic h, the p-value for the higher criticism test statistic given in (2.1) *is*

$$\mathrm{pr}\left(HC \geq h\right) = 1 - \prod_{k=1}^{d} \sum_{a=0}^{d-k} q_{k,a}. \tag{2.8}$$

Proof. The result follows immediately by the definition of $q_{k,a}$ combined with Theorem 1, equation (2.3), and equation (2.6). □

Obtaining the higher criticism p-value analytically in finite samples is a three-step process. Firstly, the observed test statistic h is computed by finding the supremum in (2.1) which is done by finding the maximum value attained over all observed test statistics $|Z_i|$. Upon computing h, the partition in (2.5) is computed. Lastly, the $q_{k,a}$s are calculated using this partition. As there are $d(d+1)/2$ different $q_{k,a}$ terms, each requiring a sum of order d in order to be calculated, the total computation time for this last step is $O(d^3)$.

The p-value calculation has been implemented in the statistical computing software, R, in the package *GHC*. The precision of this method (as well as the inaccuracy of the asymptotic p-values) is confirmed by the Type I error simulations in Table 2.1. The computation time in seconds for a given d on a 2·30 GHz laptop with 4 Gb memory can be well approximated by the polynomial (3·69e-4)p - (6·98e-6)p^2 + (3·63e-6)p^3. This corresponds to 0·007 seconds, 0·45 seconds, 28·9 seconds, and 4 hours, for $d = 10, 50, 200,$ and 1000, respectively. For comparison purpose, we also present in Table 2.1 the empirical type I error rates calculated using the asymptotic distribution in (2.2). The results show they are subject to considerable bias.

2.4 Power Simulations

We compare the power of the higher criticism to competing methods through simulation in the context of genetic association studies. An n (number of individuals) by d (number of genetic variants) genotype matrix G is generated such that the rows are independent and the columns are autocorrelated with correlation parameter ρ. Marginally, each $G_{ij} \sim Binomial(2, 0{\cdot}3)$. Letting ϵ be the n by 1 vector of independent standard normal noise and β be the d by 1 vector of regression coefficients, the phenotypes are generated according to $y = G\beta + \epsilon$. The test for the association between the jth genetic variant and y is $Z_j = G_j^T(y - \bar{y})/(\hat{\sigma}^2 G_j^T G_j)^{1/2}$ where G_j is the jth column of G and $\hat{\sigma}^2$ is the sample variance.

Power is calculated for a region of size $d = 40$, with 10% and 5% sparsity with autocorrelation $\rho = 0, 0{\cdot}05, 0{\cdot}1, 0{\cdot}15, 0{\cdot}2, 0{\cdot}25, 0{\cdot}3, 0{\cdot}35$. In each setting, 1000 iterations resulted in power estimates. For each iteration, causal variants were selected at random. If the jth variant is causal, it has $\beta_j = 0{\cdot}11$ for the 5% sparsity case and $\beta_j = 0{\cdot}08$ for the 10% sparsity case. If the jth variant is non-causal then $\beta_j = 0$. In the cases where $\rho > 0$, the test statistics need to first be decorrelated as in Hall and Jin (2010) so that the higher criticism is applied instead to $U^{-1/2}Z$ (see Section 2). The sequence kernel association test of Wu et al. (2011b) is provided as a comparison alongside the standard likelihood ratio test, and the results are in Fig. 2.2.

The higher criticism has the highest power in the higher sparsity situations, while the likelihood ratio test and the sequence kernel association test have higher power for lower sparsity. The sequence kernel association test seems to benefit greatly from increased correlation, while the higher criticism loses power as ρ increases. The reason for this is because the transformation $U^{-1/2}Z$ deflates some of the signal for larger ρ, leading to lower power. Overall, it seems that all three methods can be viable. The higher criticism is a good complement to the other methods, and performs better in the presence of weak correlation among the variants and sparse signals.

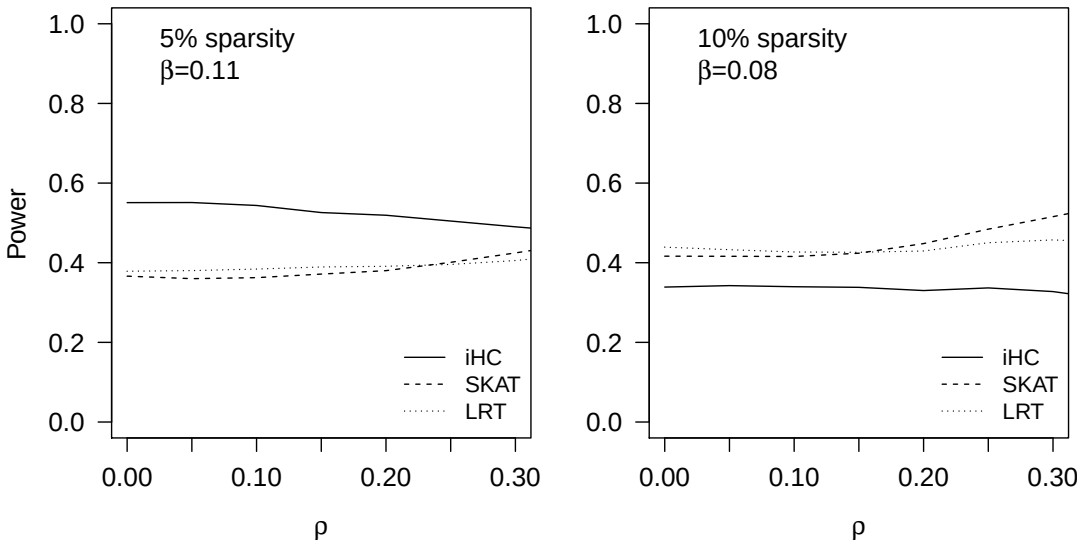

Figure 2.2: The power of the higher criticism applied to the decorrelated test statistics (iHC) is compared to the likelihood ratio test (LRT) and the sequence kernel association test (SKAT). Power is simulated for each ρ that is a nonnegative multiple of $0{\cdot}05$ and the smoothed results are displayed.

2.5 Data analysis

We apply our p-value computation of the higher criticism to a case-control lung cancer genome-wide association study conducted at the Massachusetts General Hospital, which aims at identifying genes that are associated with the risk of non-small-cell lung cancer. The study consists of a total of 984 cases and 970 controls. We use the higher criticism to test for the association of lung cancer risk and each gene, which consists of multiple genetic variants. We analyze a total of 14396 genes throughout the genome with $d = 20$ genetic variants per gene on average. The majority (92%) of genes have $d < 50$ genetic variants with the largest d being 1665.

For each gene, we calculate the marginal test statistics Z_j $(j = 1, \cdots, d)$ for genetic variant j by fitting a logistic regression model of case-control status on the genetic variant j while controlling for age, sex, smoking, and principal components to control for population stratification (Price et al., 2006). As done in Section 3, the marginal test statistics corresponding to the genetic variants within the same gene are decorrelated using the Hall and Jin (2010) approach.

Three of the top most significant genes in this analysis were CHRNA5, MYH10,

and CLPTM1L, and they each have been independently found to be associated with lung cancer in previous studies (Spitz et al., 2008; Zienolddiny et al., 2009; Wang et al., 2012). For the CHRNA5, MYH10, and CLPTM1L genes, the higher criticism p-values are 6·37e-4, 6·42e-4, and 1·29e-2; the sequence kernel association test p-values are 8·19e-4, 0·41, and 1·54e-5; and the likelihood ratio test p-values are 4·92e-3, 1·96e-3, and 7·05e-4, respectively.

As observed in the power simulations of Section 3, the higher criticism test and the sequence kernel association test complement each other for these most significant genes. In order to correct for multiple comparisons, the higher criticism can be used for signal identification in a hierarchical fashion on the 14396 p-values (Donoho and Jin, 2008) A threshold is set at the point where the test statistics attain the supremum for the higher criticism and all genes with test statistics more extreme than that threshold are declared to be disease-associated. This procedure controls for the false non-discovery rate (Ahdesmäki et al., 2010). In our case this procedure selected the four most significant genes.

As a clear example of why the asymptotic distribution of the higher criticism can be wildly inaccurate for finite d settings, the asymptotic p-values for CHRNA5, MYH10, and CLPTM1L are 5·97e-72, 3·83e-64, and 3·27e-12, respectively. These inaccuracies are amplified in the tails of the distribution which is why the asymptotic p-values differ by so many orders of magnitude from the exact p-values obtained from Theorem 2.

2.6 Discussion

The proposed analytic method for calculating the p-value of the higher criticism is exact for any arbitrary finite d for normally distributed test statistics. It is computationally fast for the common non-large d settings encountered in gene-level analysis in genome-wide association studies. For non-normally distributed outcomes, such as binary outcomes in case-control studies, accuracy of the proposed calculations depends on the accuracy of the normality approximation of individual marginal test statistics. For large sample sizes, which is often the case in genome-wide association studies, the nor-

mality assumption of individual test statistics holds quite well and the proposed p-value calculations for an arbitrary d have high accuracy.

While the vast majority of genes in the genome will have only dozens of genetic variants, there may be a very few large genes with d being in the thousands. For these large genes, simulating the null distribution could be less of a computational burden than the analytic p-value calculation given by Theorem 2. Hence a mixture of both techniques could lead to a faster overall analysis of genome-wide association data. For example, Theorem 2 could be used to calculate p-values unless the gene has $d > 500$ in which case simulation of the null distribution could be used to obtain the p-value. In the presence of correlation, the decorrelation transformation (Hall and Jin, 2010) could dampen the non-null signals when the correlation between some of the marginal test statistics is moderate or strong. It would be of future interest to develop an alternative higher criticism method to account for the correlation more effectively to improve the test power.

The Generalized Higher Criticism for Testing SNP-sets in Genetic Association Testing

Ian Barnett, Rajarshi Mukherjee and Xihong Lin

Department of Biostatistics

Harvard School of Public Health

3.1 Introduction

With the abundance of genome-wide association studies (GWAS) and with the increase in large-scale sequencing studies, so too must there be development of methodology capable of testing complex traits for genetic associations. Initial analysis of GWAS data has shown to be insufficient in explaining much of the heritability of complex non-mendelian traits, suggesting that there may not be enough power to simply test single nucleotide polymorphisms (SNPs) for marginal associations (Manolio et al., 2009). Due to single SNP analysis being underpowered, particularly for low frequency SNPs, region-based analyses have become more popular in genetic association studies (Li and Leal, 2008). Genes, gene networks, and pathways are examples of SNP-sets that have sparse subsets of SNPs that can contribute to disease risk. Methodology that does not require strong marginal SNP effects but is capable of aggregating these small and sparse SNP effects together into a detectable signal is needed to help find the causes of the missing heritability.

Hunter et al. (2007) analyzed the Cancer Genetic Markers of Susceptibility (CGEM) GWAS breast cancer data, a case-control study with post-menopausal women of European ancestry, and looked at the effects of individual SNPs across the genome. In an effort to gain more power by looking at SNP-sets instead of individual SNPs, Wu et al. (2010) scanned the genome at the gene level and found the FGFR2 region to be significantly associated with breast cancer. Though the region contained 35 SNPs, the signal was sparse with only four of the SNPs showed any association with disease. Marginally, none of those four SNPs were at genome-wide significance levels due to the large multiple testing problem. However, when the signal from all SNPs in the region are were treated as a unit, then the association became more significant. Due to their close proximity, it is common for SNPs within the same gene to be correlated, also known as being in linkage disequilibrium (LD). This demonstrates the importance of having methodology for testing SNP-sets that is powerful when the signal is sparse while accounting for the LD between SNPs.

The higher criticism (HC) is a SNP-set test that combines information over all the

marginal test statistics (Donoho and Jin, 2004) and is ideal for detecting a sparse few disease-associated SNPs out of a much larger pool of unassociated SNPs than comparable methods (Arias-Castro et al., 2011). While originally designed for high dimensional settings with independence between SNPs, an adaptation to the higher criticism, the innovated higher criticism (iHC), first transforms the test statistics to independent test statistics using the Cholesky decomposition of the correlation matrix and then applys the higher criticism after the transformation (Hall and Jin, 2010). This transformation can be unstable when SNPs are in high LD. Our numerical results also suggest that the HC based on the transformed test statistics is subject to considerable loss of power in the presence of correlation among the SNPs. Additionally, as SNP-sets often are not large and these implementations of the higher criticism give only asymptotic results, they cannot be directly applied to most SNP-sets without the use of simulation.

Several methods have been proposed for SNP-set testing, such as MinP and variance-component tests. MinP calculates the marginal test statistic for each SNP in the SNP-set and then uses the maximum (or most extreme) marginal test statistics as the representative test statistic while adjusting for correlation within the SNP-set (Conneely and Boehnke, 2007; Moskvina and Schmidt, 2008; Cheverud, 2001; Nyholt, 2004; Kimmel and Shamir, 2006; Han et al., 2009; Zhang and Liu, 2011). MinP has low power when a sparse set of individual SNPs don't have strong signals and instead combine together to form a strong signal. Variance-component tests such as the Sequence Kernel Association Test (SKAT) (Wu et al., 2011b) offer an alternative to MinP for detecting SNP-set associations by combining all SNP information over the SNP-set (Liu et al., 2007, 2008; Lin, 1997; Neale et al., 2011). However if the signal in the region is sparse, then SKAT can have low power due to giving equal weight to noncausal SNPs in the region, which can cover up the signal with noise. As a result, the limitation for both MinP and SKAT is that they both may not be as effective as the higher criticism in accounting for sparse alternatives (Arias-Castro et al., 2011).

In this paper we present the generalized higher criticism (GHC) test statistic that is suitable for SNP-sets containing a finite set of correlated markers that does not require any transformation of the original test statistics. In contrast to the original higher criticism, the

GHC is flexible to any correlation structure and its analytic p-values can be obtained in an accurate and computationally efficient manner. The power of GHC relative to iHC, SKAT, and MinP is compared over extensive simulations using SNP-sets with varied correlation structures. While MinP and SKAT are sensitive to sparsity, the robust power of GHC is demonstrated through simulation over regions with varying degrees of sparsity. The robust nature of GHC is also seen when these methods are compared in their analysis of the CGEM GWAS breast cancer data.

The remainder of the paper is organized as follows. In Section 2, we introduce the SNP-set generalized linear model. In Section 3, we present the asymptotic properties of the HC as well as the problems it has accounting for correlation in SNP-sets. In Section 4, we describe the procedure for obtaining p-values for the GHC. In Section 5, the asymptotic detection boundary for GHC is established. In Section 6, we evaluate the performance of the GHC relative to comparable methods using simulation. In Section 7, the GHC and competing methods are used to analyze the CGEM breast cancer GWAS data. Finally, we conclude with discussions in Section 8.

3.2 Generalized linear model and marginal SNP score test statistics

We consider a sample of N individuals genotyped over a region with p observed SNPs in a SNP-set. Possible SNP-sets include genes, gene networks, or genetic pathways. Individuals have phenotypes $\boldsymbol{Y} = [Y_1, ..., Y_N]^T$. The N by p genotype matrix $\boldsymbol{G}$ is constructed such that $\boldsymbol{G_i} = [G_{i1}, ..., G_{ip}]^T$ with $\boldsymbol{G_i^T}$ as the ith row vector of $\boldsymbol{G}$ containing the genotypes for the ith individual. The N by q matrix $\boldsymbol{X}$ contains non-genetic covariate information for the sample with $\boldsymbol{X_i} = [X_{i1}, ..., X_{iq}]^T$ and $\boldsymbol{X_i^T}$ being the ith row vector of $\boldsymbol{X}$ containing the covariate values for the ith individual. We suppose that conditional on $(\boldsymbol{X}_i, \boldsymbol{G}_i)$, Y_i follows a distribution in the exponential family (MacCullagh and Nelder, 1989) $f(Y_i) = exp\{(Y_i\theta_i - b(\theta_i))/a_i(\phi) + c(Y_i, \phi)\}$, where $f(Y_i)$ is the conditional distribution of $Y_i|(\boldsymbol{X}_i, \boldsymbol{G}_i)$, $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions, θ_i is the canonical parameter, and ϕ is the dispersion parameter. In order to construct a marginal test between the jth SNP

and $\boldsymbol{Y}$ we model $\mu_i = E(Y_i|\boldsymbol{G_i}, \boldsymbol{X_i}) = b'(\theta_i)$ as using the GLM (MacCullagh and Nelder, 1989)

$$g(\mu_i) = \boldsymbol{X_i}^T\boldsymbol{\alpha} + \boldsymbol{G_i}^T\boldsymbol{\beta} \tag{3.1}$$

where $g(\cdot)$ is a link function and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are the regression coefficients. For simplicity, we here restrict to the canonical link. The variance of Y_i is $Var(Y_i) = a_i\phi v(\mu_i)$, where $v(\mu_i) = a_i\phi b''(\theta_i)$ is a variance function. We are interested in testing for the overall effect of the SNP set $\boldsymbol{G}_i$, which corresponds to the global null $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$.

Letting $\boldsymbol{W} = diag\{a_1\phi v(\hat{\mu}_{01}), ..., a_n\phi v(\hat{\mu}_{0n})\}$ and $\boldsymbol{P} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}$, the marginal score test statistic for β_j under the global null is

$$Z_j = \frac{\boldsymbol{G_j}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_0)}{\sqrt{\boldsymbol{G_j}^T\boldsymbol{P}\boldsymbol{G_j}}} \tag{3.2}$$

where $\hat{\boldsymbol{\mu}}_0 = \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\alpha}})$, and $\hat{\boldsymbol{\alpha}}$ is the MLE of $\boldsymbol{\alpha}$ under the null model of $g(\mu_i) = \boldsymbol{X_i}^T\boldsymbol{\alpha}$. These individual SNP test statistics are asymptotically jointly distributed as $\boldsymbol{Z} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$ where we estimate $Cov(Z_j, Z_k) = \boldsymbol{\Sigma}_{jk}$ by

$$\hat{\boldsymbol{\Sigma}}_{jk} = \frac{\boldsymbol{G_j}^T\boldsymbol{P}\boldsymbol{G_k}}{\sqrt{\boldsymbol{G_j}^T\boldsymbol{P}\boldsymbol{G_j}}\sqrt{\boldsymbol{G_k}^T\boldsymbol{P}\boldsymbol{G_k}}}$$

While the $\boldsymbol{Z}$ are correlated we define the uncorrelated transformed test statistics $\boldsymbol{Z}^*$ to be

$$\boldsymbol{Z}^* = \boldsymbol{U}^{-1}\boldsymbol{Z} \sim MVN(\boldsymbol{0}, \boldsymbol{I}_p)$$

where $\boldsymbol{U}\boldsymbol{U}^T = \hat{\boldsymbol{\Sigma}}$ is the Cholesky decomposition.

3.3 The higher criticism

The higher criticism tests $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ from model (3.1) against the alternative that a sparse set of the β_j are nonzero. An idea proposed first in passing by Tukey, the higher criticism was developed by Donoho and Jin (2004) for summary statistics in the setting where the under the alternative the marginal test statistics come from a mixture of normal

random variables, as well as by Arias-Castro et al. (2011) in the regression setting. Because the $\boldsymbol{Z}$ are correlated, Hall and Jin (2010) proposed a higher criticism test based on the transformed $\boldsymbol{Z}^*$. To do so let

$$S^*(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j^*| \geq t\}}$$

Note that under H_0, $S^*(t) \sim Binomial(p, 2\bar{\Phi}(t))$ where $\bar{\Phi}(t) = 1 - \Phi(t)$ is the survival function of the normal distribution. The innovated higher criticism test statistic is defined as

$$iHC = sup_{t \geq t_0} \left\{ \frac{S^*(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \right\}$$

for some $t_0 \geq 0$. This test rejects H_0 for large values of iHC. If we allow $p \to \infty$ the higher criticism has been shown to be powerful for high sparsity situations (i.e. when the number of $\beta_j \neq 0$ are less than $\sqrt{p}$) (Donoho and Jin, 2004; Hall and Jin, 2010). For very large p, iHC can be viewed as the supremum of a normalized empirical process which converges asymptotically to a Gumbel distribution at a very slow rate of $O\{(\log p)^{-1/2}\}$ (Jaeschke, 1979). With such a slow rate of convergence, the size of the test is drastically incorrect when using the asymptotic distribution to calculate P-values for p as large as a million (Barnett and Lin, 2013). Around 92% of genes in GWAS have $p < 50$, and with most functional networks and pathways containing a few hundred genes, the size of SNP-sets in both gene-level and pathway-level analyses are generally not large enough for the asymptotic distribution of the iHC to be of any practical use.

For finite p we must take a different analytic approach to find the distribution of the iHC statistic. An exact P-value calculation for iHC that does not rely on asymptotics has been developed for this situation (Barnett and Lin, 2013). This P-value calculation is an exact and computationally efficient method for all finite p, and therefore ideal for testing SNP-sets in genetic association studies. However, there are drawbacks due to having to first transform the marginal test statistics from $\boldsymbol{Z}$ into $\boldsymbol{Z}^*$. In the presence of even moderately small correlation, there can be a significant loss of power due to the noise diluting the sparse signals after being mixed in the transformation. For example, alleles in the FGFR2 gene have been linked to breast cancer risk (Hunter et al., 2007). The CGEM breast cancer data has 35 SNPs in the gene, four of which have marginal

test statistics greater than 4.3 in absolute value indicating a strong association between FGFR2 and breast cancer incidence. However, after transforming these test statistics to become $\boldsymbol{Z}^*$, none of the transformed statistics exceed 2.6 in absolute value (Figure 3.1). These transformed test statistics are so attenuated towards the null that it can lead to a significant loss in power. This motivates our generalization of the higher criticism to accomodate the use of the original untransformed correlated test statistics $\boldsymbol{Z}$ in order to avoid this loss of power.

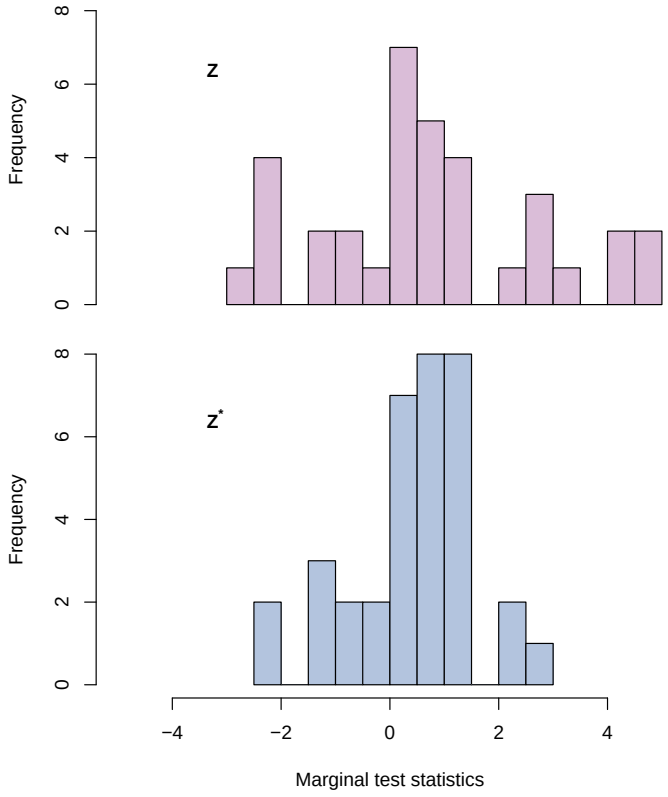

Figure 3.1: The marginal test statistics for 35 SNPs from the FGFR2 gene, each with $MAF > 0.05$, from the CGEM genetic association study of breast cancer are plotted. The original test statistics $\boldsymbol{Z}$ are in the top histogram, while the transformed test statistics $\boldsymbol{Z}^* = \boldsymbol{U}^{-1}\boldsymbol{Z}$ are in the bottom histogram.

3.4 The generalized higher criticism

If the LD structure in a gene or SNP-set is even moderately weak, it is likely that transforming $\boldsymbol{Z}$ by $\boldsymbol{U}^{-1}$ can result in the transformed test statistics $\boldsymbol{Z}^*$ being underpowered. In addition, for stronger LD, the matrix inverse operation of $\boldsymbol{U}^{-1}$ can be quite unstable. In order to avoid this problematic transformation we propose the generalized higher criticism test statistic based on the original $\boldsymbol{Z}$.

3.4.1 Definition of the generalized higher criticism

We define $S(t)$ as

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}.$$

For general $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$, $S(t)$ is no longer binomial. The correlation, whether positive or negative, increases the variability of $S(t)$. To see this we estimate $Cov(S(t_j), S(t_k))$ using the sample correlation $\hat{\boldsymbol{\Sigma}}$ and the following identity:

Theorem 3. Let $\overline{r^n} = \frac{2}{p(1-p)} \sum_{1 \leq k < l \leq p} (\Sigma_{kl})^n$ and let $\mathcal{H}_i(t)$ be the Hermite polynomials: $\mathcal{H}_0(t) = 1$, $\mathcal{H}_1(t) = t$, $\mathcal{H}_2(t) = t^2 - 1$ and so on. Then

$$\begin{aligned} Cov\left(S(t_k), S(t_j)\right) &= p[2\bar{\Phi}(max\{t_j, t_k\}) - 4\bar{\Phi}(t_j)\bar{\Phi}(t_k)] \\ &\quad + 4p(p-1)\phi(t_j)\phi(t_k) \sum_{i=1}^{\infty} \frac{\mathcal{H}_{2i-1}(t_j)\mathcal{H}_{2i-1}(t_k)\overline{r^{2i}}}{(2i)!} \end{aligned}$$

The proof of Theorem 3 is left to Appendix C.1. Using Theorem 3 with $\hat{\boldsymbol{\Sigma}}$ instead of $\boldsymbol{\Sigma}$ we obtain estimates of $Var(S(t))$. Though Theorem 3 involves an infinite sum, the terms tend to zero so rapidly that in practice we suggest that only the first 8 terms are necessary for estimating the covariance with great accuracy. With these estimates we define the generalized higher criticism test statistic to be:

$$GHC = \sup_{t \geq t_0} \left\{ \frac{S(t) - p \cdot 2\bar{\Phi}(t)}{\sqrt{\widehat{Var}(S(t))}} \right\} \tag{3.3}$$

For simplicity we will assume $t_0 = 0$. In the independent case when $\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}_p$ the GHC

statistic reduces to HC. The stronger the correlation structure, the larger the denominator (3.3) becomes. However the GHC numerator tends to be larger than the HC numerator due to the transformed $\boldsymbol{Z}^*$ being attenuated towards the null.

The asymptotic properties of GHC are highly dependent on the correlation $\boldsymbol{\Sigma}$. If $\Sigma_{jk} = 0$ for $|j-k| > b$ for some fixed bandwidth b, then as $N, p \to \infty$ Hoeffding and Robbins (1948) show that $(S(t) - p \cdot 2\bar{\Phi}(t))(\widehat{Var}(S(t)))^{-1/2}$ tends toward the standard normal distribution with their central limit theorem for dependent random variables. By taking the supremum over all $t \geq t_0$ we are then faced with a gaussian process, and the same convergence issues seen with the distribution of iHC are present for GHC as well. In order to obtain accurate GHC P-values, we propose in the next section a P-value calculation for GHC in finite p settings that does not rely on asymptotics.

3.4.2 Calculation of the generalized higher criticism P-value

For a given observed GHC statistic, h, we show in Appendix **A.2** that the corresponding P-value is

$$pr\left(GHC \geq h\right) = 1 - \prod_{k=1}^{p} \sum_{a=0}^{p-k} q_{k,a} \tag{3.4}$$

where for $k > 1$

$$q_{k,a} = \sum_{m=0}^{p-k+1} pr\left(S(t_k) = a \middle| S(t_{k-1}) = m, \bigcap_{l=1}^{k-2} \{S(t_l) \leq p - l\} \right) \frac{q_{k-1,m}}{\sum_{l=0}^{p-k+1} q_{k-1,l}},$$

for $k = 1$, $q_{1,a} = pr(S(t_1) = a)$, and t_k is solved for in the equation

$$h\sqrt{\widehat{Var}(S(t_k))} + 2p\bar{\Phi}(t_k) = p - k + 1 \tag{3.5}$$

for each $k \in \{1, ..., p\}$.

When the test statistics are independent $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, then $S(t)$ is the sum of independent indicator variables and the distribution of $S(t_k)$ conditional on $S(t_{k-1}) = m$ and $\cap_{l=1}^{k-2}\{S(t_l) \leq p - l\}$ is binomial with m events and probability of success $\bar{\Phi}(t_k)/\bar{\Phi}(t_{k-1})$. When calculating the P-value for GHC for general $\boldsymbol{\Sigma}$, $S(t)$ is not binomially distributed. Instead, in order to account for the additional variability the distribution of $S(t_k)$ con-

ditional on $S(t_{k-1}) = m$ and $\cap_{l=1}^{k-2}\{S(t_l) \leq p - l\}$ is approximated with beta-binomial distribution.

The condition distribution of $S(t_k)$ is approximated by

$$pr\left(S(t_k) = a \middle| S(t_{k-1}) = m, \bigcap_{l=1}^{k-2}\{S(t_l) \leq p - l\}\right) \approx pr\left(S(t_k) = a \mid S(t_{k-1}) = m\right).$$

This approximation is an equality if the marginal test statistics are independent due to the markov property of empirical processes (Gaenssler, 1983). The variance of $S(t_k)$ conditional on $S(t_{k-1}) = m$ is obtained by conditioning on which of the m different $|Z_j|$ are greater than t_{k-1}. The expectation is $m \cdot \bar{\Phi}(t_k)/\bar{\Phi}(t_{k-1})$ just like in the independent case. Using these first two moments, the parameters of the beta-binomial distribution (α and β) are solved for numerically using moment matching in the equations

$$m\frac{\bar{\Phi}(t_k)}{\bar{\Phi}(t_{k-1})} = \frac{m\alpha}{\alpha+\beta}$$

$$2\frac{\binom{m}{2}}{\binom{p}{2}}\sum_{j<l}\frac{pr(|Z_j|, |Z_l| > t_k)}{pr(|Z_j|, |Z_l| > t_{k-1})} + m\frac{\bar{\Phi}(t_k)}{\bar{\Phi}(t_{k-1})} - \left(m\frac{\bar{\Phi}(t_k)}{\bar{\Phi}(t_{k-1})}\right)^2 = \frac{m\alpha\beta(\alpha+\beta+m)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where $pr(|Z_j|, |Z_l| > t_k)$ can be obtained as in Schwartzman and Lin (2011) (see Appendix **A.1**)). The unconditional distribution of $S(t_k)$ can be obtained in the same way by substituting $t_{k-1} = 0$ and $m = p$.

3.5 The detection boundary of GHC

Though the GHC test is designed for testing SNP-sets where p is often not large, its asymptotic properties can still help provide some insight into its performance. One way to analyze the validity of a testing procedure in an asymptotic sense in multivariate problems is to study the minimax decision errors. Although minimax considerations are often only of theoretical interest and involves a pessimistic view towards error measurements, it provides an optimality criteria in multivariate problems where most powerful tests, even in the class of unbiased tests, do not exist. In the context of Gaussian linear regression, the minimax detection boundary of testing global null against sparse alternatives has been extensively studied Arias-Castro et al. (2011); Ingster et al. (2010). Extend-

ing these results to the generalized linear model framework is a more difficult problem (Mukherjee et al., 2013) and so we will consider the case where $\boldsymbol{\mu} = \boldsymbol{G}\boldsymbol{\beta}$. We will show that the GHC test attains the same detection boundaries obtained in (Arias-Castro et al., 2011; Ingster et al., 2010). In the following, we first state our problem and asymptotic framework properly and then discuss optimality properties of GHC.

Let $M(\boldsymbol{\beta}) = \sum_{j=1}^{p} I(\beta_j \neq 0)$ and let $R_k^p = \{\boldsymbol{\beta} \in \mathbb{R}^p : M(\boldsymbol{\beta}) \geq k\}$. For some $A > 0$, we are interested in testing the global null hypothesis

$$H_0 : \boldsymbol{\beta} = 0 \text{ vs } H_1 : \boldsymbol{\beta} \in \Theta_k^A = \{\boldsymbol{\beta} \in \bigcup_{k' \geq k} R_{k'}^p : \min\{|\beta_j| : \beta_j \neq 0\} \geq A\}. \tag{3.6}$$

Set $k = p^{1-\alpha}$ with $\alpha \in (0, 1]$. We note that this types of alternatives has been considered by Arias-Castro et al. (2011), referred to as the "*Sparse Fixed Effects Model*" or SFEM.

We study the asymptotic properties of our testing problem in the high-dimensional regime, *i.e.*, with $p \to \infty$ and $n = n(p) \to \infty$. To introduce our minimax formulation of the problem, let for any test T the worst case risk be

$$R(\mathsf{T}) := \mathbb{P}_0(\mathsf{T} = 1) + \max_{\boldsymbol{\beta} \in H_1} [\mathbb{P}_{\boldsymbol{\beta}}(\mathsf{T} = 0)] \,. \tag{3.7}$$

where a test is a measurable function of the data taking values in $\{0, 1\}$. Adopting terminology from Arias-Castro et al. (2011), we say that a sequence of tests $\{\mathsf{T}_{n,p}\}$ is *asymptotically powerful* if $\lim_{p\to\infty} R(\mathsf{T}_p) = 0$ and we say that it is *asymptotically powerless* if $\liminf_{p\to\infty} R(\mathsf{T}_{,p}) = 1$. The detection boundary of the testing problem (3.6) is the demarcation of signal strength A which determines whether all tests are asymptotically powerless (we call this Lower Bound of the problem) or there exists some test which is asymptotically powerful (we call this the Upper Bound of the problem).

Arias-Castro et al. (2011) demonstrate that under certain low-correlation conditions on $\boldsymbol{\Sigma}$, the higher criticism is optimally asymptotically powerful. For $\alpha > \frac{1}{2}$, Arias-Castro et al. (2011) demonstrate the detection boundary

$$\rho^*(\alpha) = \begin{cases} \alpha - 1/2 & \text{if } 1/2 < \alpha < 3/4 \\ (1 - \sqrt{1-\alpha})^2 & \text{if } 3/4 \leq \alpha < 1 \end{cases} \tag{3.8}$$

is such that all tests are asymptotically powerless if $A = \sqrt{2r\log p}$ with $r < \rho^*(\alpha)$. In Theorem 3 of Arias-Castro et al. (2011) they show that a test based on

$$HC^*(s) = \max_{t \in [s, \sqrt{5\log p}] \cap \mathbb{N}} \left\{ \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \right\}$$

is asymptotically powerful everywhere above the detection boundary for any given α and $r > \rho^*(\alpha)$ when $s = \sqrt{2\min(1, 4\rho^*(\alpha))\log p}$ and when the correlation structure is not too strong. More formally, this holds under the following two conditions on $\mathbf{\Sigma}$:

(i) **Strong correlation condition**: $|\Sigma_{jk}| < 1 - (\log p)^{-1}$ for every $j \neq k$

(ii) **Weak correlation condition**: For all j, $|\{k : |\Sigma_{jk}| > \gamma\}| \leq \Delta$

where $\Delta = O(p^\epsilon)$, $\gamma^2 p^{1-\alpha}(\log p)^3 \to 0$, and $\gamma^3 = O(p^{\epsilon + 5\alpha - 4})$ for all $\epsilon > 0$. The strong correlation condition does not allow any pairwise correlations to be too large, while the weak correlation condition restricts the number of pairwise correlation much different from zero that each SNP can have. This correlation framework fits well into the genetic context, where regions with large p will tend to have an upper limit on the number of pairwise correlations far from zero for each SNP due to the fact that LD tends to decrease as the distance between SNPs increases.

The difference between $HC^*(s)$ and GHC is that $HC^*(s)$ assumes $S(t)$ to be binomially distributed and normalizes by the binomial variance even in the presence of correlation whereas GHC normalizes by the correct variance term that takes correlation into account. We will show that, despite this difference, GHC is asymptotically powerful in the same regions as $HC^*(s)$. Though GHC was defined in (3.3) to take the supremum over $t > t_0$ for some fixed t_0, this only makes sense in the finite p setting. Asymptotically, we will instead take the supremum over $[s, \sqrt{5\log p}] \cap \mathbb{N}$ as done with $HC^*(s)$.

Theorem 4. Under the conditions (i) and (ii) on $\mathbf{\Sigma}$ then the test based on GHC with the supremum taken over $[\sqrt{2\min(1, 4\rho^*(\alpha))\log p}, \sqrt{5\log p}] \cap \mathbb{N}$ is asymptotically powerful against alternatives defined by sparsity $p^{1-\alpha'}$, $\alpha' \geq \alpha > 1/2$, $A = \sqrt{2r\log p}$, and $r > \rho^*(\alpha')$

The proof of Theorem 4 is left to Appendix C.2. The implications of Theorem 4 are that when correlation between SNPs is fairly weak, then GHC has the same asymp-

totic performance as HC. It also suggests that the GHC test can have greater power than comparable methods when there are a sparse few SNPs associated with disease in large p settings. However it is important to note that this large p performance will not necessarily translate to finite p situations with stronger correlation structures that are frequently encountered in genetic association studies. Next, we evaluate this finite p performance through simulation.

3.6 Simulation studies

3.6.1 Type I error of GHC

In order to determine the accuracy of the P-value calculation for the GHC, the Type I error of the method is estimated through simulation. To mimic the CGEM breast cancer data, a subset of 35 common HapMap SNPs with minor allele frequency (MAF) greater than 0.05 in the FGFR2 gene were simulated using the LD structure from the CEU population in the HapMap project. Because the Type I error is possibly dependent on the strength of the LD in the region being tested, two subsets of FGFR2 are separately considered. A subset of 8 SNPs in high LD with each other and a different subset of 8 SNPS in low LD with each other are used to estimate the Type I error in high and low LD cases, respectively.

For each subset, the 8 SNPs are treated as a SNP-set and 1000 cases and 1000 controls are generated from logistic regression model (3.1) with $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{X}_i = 1$, and $\boldsymbol{\alpha} = -0.5$.The GHC P-value in (3.4) is calculated for each simulated dataset and this is repeated 50 million times in order to have Type I error estimates for genome-wide significance levels as small as 10^{-6}. To see that the size of the GHC is correct in the more general case beyond FGFR2, we also simulated Type I error in the same way except by using randomly selected genes from chromosome 5 for each iteration. In each setting, Type I error is accurate at all significance levels, though slightly conservative for stronger correlation structures (Table 3.1).

GHC Type I error simulations

Significance level	Strong LD (FGFR2 subset)	Weak LD (FGFR2 subset)	Random chr5 genes
$\alpha = 5.0 \cdot 10^{-2}$	$4.62 \cdot 10^{-2}$	$4.91 \cdot 10^{-2}$	$4.64 \cdot 10^{-2}$
$\alpha = 1.0 \cdot 10^{-2}$	$9.59 \cdot 10^{-3}$	$1.04 \cdot 10^{-2}$	$9.53 \cdot 10^{-3}$
$\alpha = 1.0 \cdot 10^{-3}$	$9.63 \cdot 10^{-4}$	$1.08 \cdot 10^{-3}$	$9.80 \cdot 10^{-4}$
$\alpha = 1.0 \cdot 10^{-4}$	$9.51 \cdot 10^{-5}$	$1.00 \cdot 10^{-4}$	$9.63 \cdot 10^{-5}$
$\alpha = 1.0 \cdot 10^{-5}$	$9.70 \cdot 10^{-6}$	$9.50 \cdot 10^{-6}$	$8.05 \cdot 10^{-6}$
$\alpha = 1.0 \cdot 10^{-6}$	$8.60 \cdot 10^{-7}$	$7.00 \cdot 10^{-7}$	$8.63 \cdot 10^{-7}$

Table 3.1: Type I error is estimated in each setting with 50 million simulations. The strong LD setting is based on a subset of 8 FGFR2 HapMap SNPs that are in high LD with one another. The weak LD setting is based on a subset of 8 FGFR2 HapMap SNPs that are in weak LD with one another.

3.6.2 Power comparisons for different LD and sparsity settings

The power of GHC, iHC, SKAT, and MinP is compared in situations where the sparsity and LD structure of the SNP-set vary. SKAT is a variance component score test that rejects the null hypothesis of $\boldsymbol{\beta} = \mathbf{0}$ for large values of the quadratic form $(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})'\boldsymbol{G}\boldsymbol{R}_\tau\boldsymbol{G}'(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$ where $\boldsymbol{R}_\tau = (1-\tau)\boldsymbol{I} + \tau\mathbf{1}\mathbf{1}'$, $\hat{\boldsymbol{\mu}}$ is the expectation of $\boldsymbol{Y}$ under the null hypothesis, and τ is selected to minimize the P-value. The package *SKAT* in the statistical computing software, R, is used to calculate the P-values. The MinP test statistic is the largest (in absolute value) marginal test statistic Z_j from (3.2) taken over all the SNPs in the region. For each setting in the power simulations, the MinP test statistic is also simulated 5000 times assuming the null distribution so that P-values can be obtained by comparing to the null distribution.

Genotype matrices were generated with $N = 2000$ (1000 cases and 1000 controls) and $p = 40$ with each SNP having MAF $= 0.30$. The genotype matrices are generated such that all causal variants have pairwise correlation ρ_1 with each other, all non-causal variants have pairwise correlation ρ_3 with each other, and each causal variant has pairwise correlation of ρ_2 with each non-causal variant. Power is simulated for $\rho_1 = 0, 0.4$, $\rho_3 = 0, 0.4$, and for ρ_2 for all non-negative multiples of 0.01 that result in positive definite $\boldsymbol{\Sigma}$. Two sparsity settings are considered, $2(5\%)$ causal variants and $4(10\%)$ causal variants, and in each case the causal SNPs are given an effect size of $\beta = 0.18$ and $\beta = 0.10$, respec-

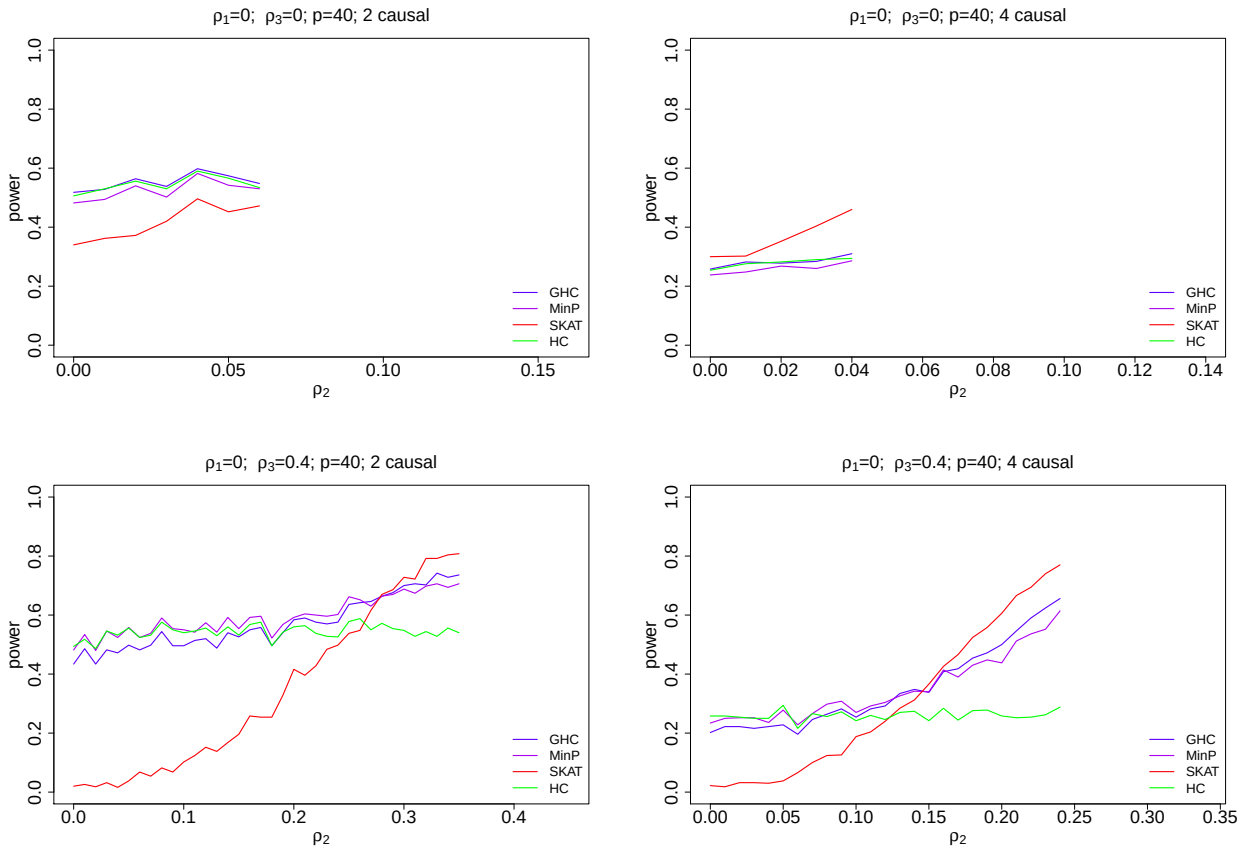

Figure 3.2: Genotypes are simulated with no correlation within the causal variants ($\rho_1 = 0$). Starting with $\rho_2 = 0$, power is estimated from 500 simulations for each possible $\rho_2 > 0$ that is a multiple of 0.01. There is a limit on how large ρ_2 can be relative to ρ_1 and ρ_3 so that the correlation matrix remains positive definite, and for this reason the range of ρ_2 values that power is estimated for varies with ρ_1 and ρ_3.

tively. Noncausal SNPs have $\beta = 0$. The dichotomous traits are generated according to

$$logit(P(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{G}_i)) = -1.8 + 0.05X_{i1} + 0.01X_{i2} + \sum_{j=1}^{p} \beta_j G_{ij} \tag{3.9}$$

where X_{i1} and X_{i2} are independent standard normal random variables. Cases and controls are generated in this fashion until 1000 cases and 1000 controls are obtained. In each sparsity and correlation setting, 500 simulations are performed and power is reported at the 0.01 significance level. The results are displayed for $\rho_1 = 0$ in Figure 3.2 and for $\rho_1 = 0.4$ in Figure 3.3.

In order to compare power in a more realistic setting with more complex LD struc-

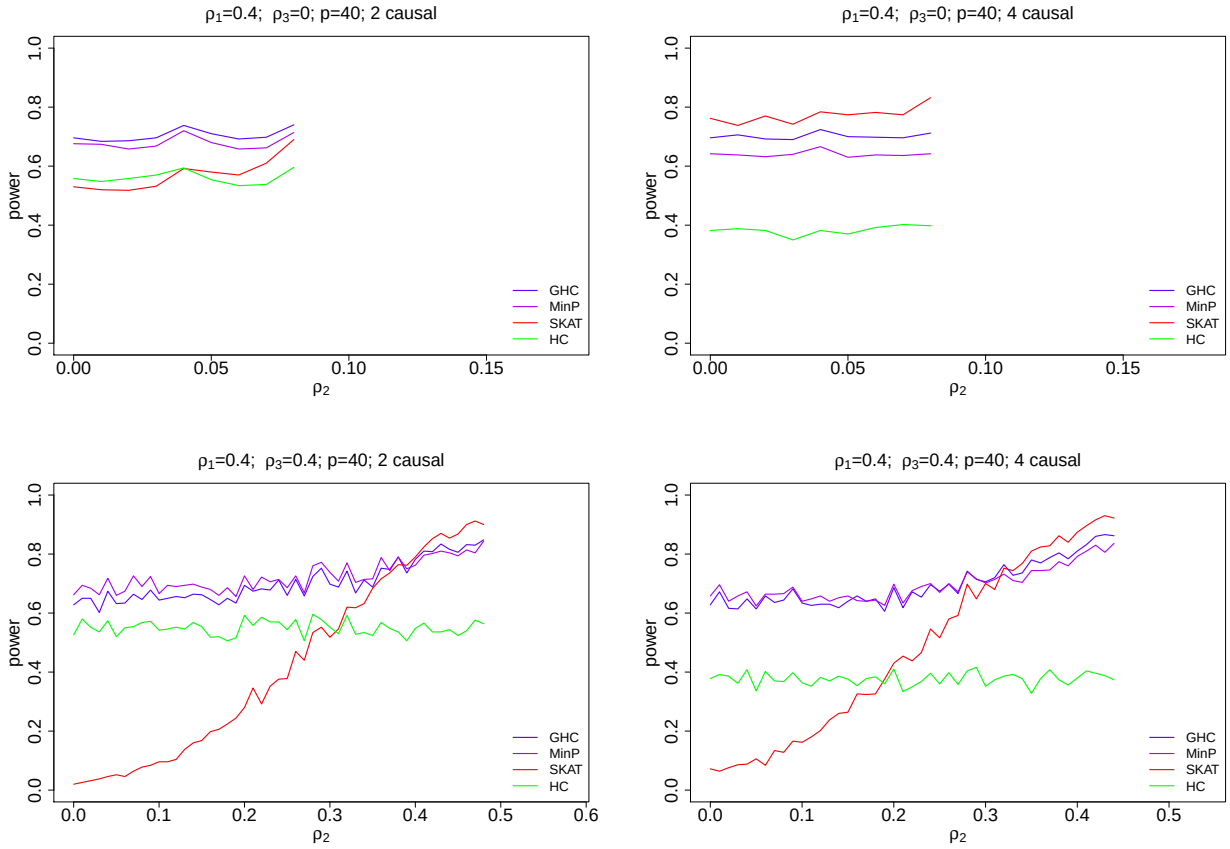

Figure 3.3: Genotypes are simulated with 0.4 correlation within the causal variants ($\rho_1 = 0.4$). Starting with $\rho_2 = 0$, power is estimated from 500 simulations for each possible $\rho_2 > 0$ that is a multiple of 0.01. There is a limit on how large ρ_2 can be relative to ρ_1 and ρ_3 so that the correlation matrix remains positive definite, and for this reason the range of ρ_2 values that power is estimated for varies with ρ_1 and ρ_3.

tures, power simulations are repeated on randomly selected genes from chromosome 5 using the 1000 cases and 1000 controls generated in the same fashion as (3.9). Genotype data is generated from common HapMap SNPs using the LD structure from the CEU population in the HapMap project. Causal SNPs are selected at random from within each gene and the two sparsity settings considered are 1 causal variant and 2 causal variants, with each causal SNP given an effect size of $\beta = 0.30$ and $\beta = 0.18$, respectively. For each of the 839 genes in chromosome 5 containing more than 1 SNP, 100 simulations are used to estimate the power. Because the power will depend on the size of each gene, p, as well as the minor allele frequency of the causal SNPs, Figure 3.4 shows a smoothed

power curve in order to represent the power averaged over all genes. In this case, ρ_2 is the median pairwise correlation between causal and non-causal variants. These results mirror the results from the artificially generated genotype matrices in Figures 3.2 and 3.3 with the only difference being that power for all methods is lower for smaller ρ_2 in Figure 3.4 because in real data lower correlation is often an artifact of low allele frequency which in turn leads to low power. This trend is not present in the artificial genotypes where allele frequency is fixed for all SNPs regardless of the correlation.

Based on these results, GHC and MinP are more similar in performance to each other in most settings than they are to either SKAT or iHC. This is not surpising because they are both tests based on the extreme marginal test statistics while ignoring the less significant test statistics except for taking correlation of the region into account. On the other hand, iHC is based on the the extreme transformed marginal test statistics, which tend to be quite different, and SKAT is a weighted sum of the squares of all the marginal test statistics. GHC improves in performance relative to MinP when ρ_1 increases, ρ_2 increases, or sparsity decreases. The iHC test has lower power than GHC in all settings with correlation present. SKAT improves relative to MinP and GHC as sparsity decreases, but has very low power when ρ_2 is low while ρ_3 is large. The reason for this is because in this situation the non-causal variants dominate the region's LD structure and because they are independent of the causal variants, they almost completely mask the sparse signal. GHC and MinP are robust to this due to their reliance on only the extreme test statistics and receive only a slight penalty in power when taking the LD into account.

Overall, MinP is ideal when there is only 1 causal variant and no LD in the region, and diminishes in power relative to other methods as sparsity decreases from there. SKAT does very well in low sparsity settings, but requires the causal variants to be highly correlated with the non-causal variants ($\rho_2 \gg 0$) if sparsity is high and non-causal variants are in LD. In contrast, GHC is is very robust to all correlation structures and all sparsity levels, and outperforms iHC in the presence of correlation between SNPs in the region.

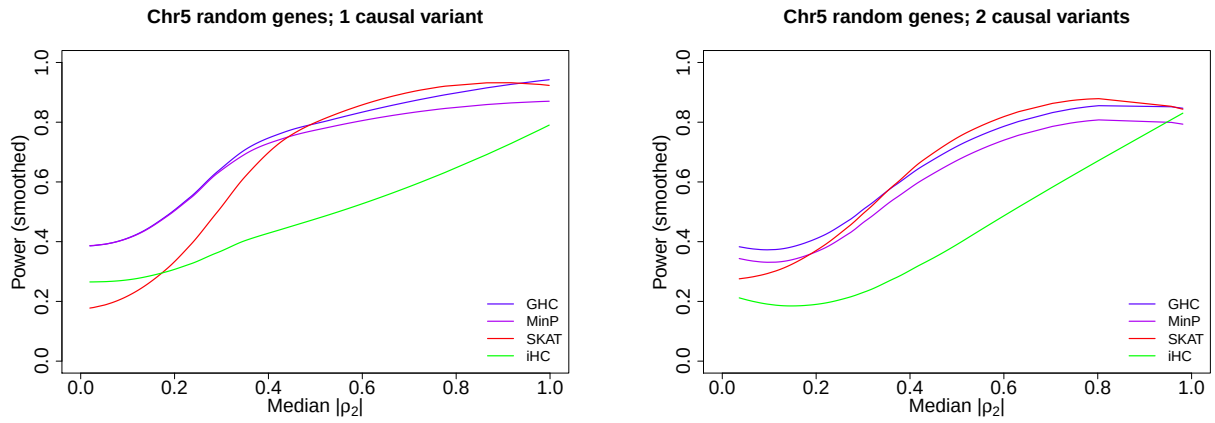

Figure 3.4: For each of the 839 genes in chromosome 5, causal SNPs are selected at random and power is estimated at the $\alpha = 0.05$ level based on 100 simulations. Additionally, the median correlation between causal SNPs and noncausal SNPs (ρ_2) is recorded. The smoothed curves to each of these power estimates is displayed.

3.7 Application to the CGEM breast cancer genetic data

The effectiveness of GHC to detect disease-associated genes next to comparable methods is explored on a breast cancer GWAS data set. A total of 1145 postmenopausal women of European ancestry with breast cancer and 1142 controls were included in the CGEM genome-wide association study (Hunter et al., 2007). These women were genotyped at 528173 loci using an Illumina HumanHap500 array. The logistic regression model (3.1) was used while controlling for the covariates: age, post-menopausal hormone usage, and the top three principal components to correct for population stratification (Price et al., 2006). Hunter et al. (2007) performed individual SNP analysis. Four SNPs in the FGFR2 region had marginal P-values less than $1.7 \cdot 10^{-5}$ and none of them were close to genome-wide significance levels. The most significantly associated SNP in FGFR2 (rs1219648) was validated in further studies. However, as we will see, by grouping information across multiple SNPs in a gene, the effect of FGFR2 is more significant. Genes with a 20kb buffer region were used to group SNPs into SNP-sets. GHC, iHC, SKAT, and MinP were all used to analyze the data for gene-level analysis, and the P-values for all four methods are displayed in Table 3.2 for genes with the most significant breast cancer

CGEM breast cancer GWAS gene-level results

Gene	p	GHC	iHC	SKAT	MinP
FGFR2	35	$2.41 \cdot 10^{-5}$	$1.07 \cdot 10^{-1}$	$3.56 \cdot 10^{-5}$	$8.05 \cdot 10^{-5}$
TBK1	11	$3.47 \cdot 10^{-4}$	$3.86 \cdot 10^{-3}$	$5.63 \cdot 10^{-5}$	$8.39 \cdot 10^{-4}$
PTCD3	12	$5.77 \cdot 10^{-5}$	$1.11 \cdot 10^{-3}$	$1.15 \cdot 10^{-4}$	$2.18 \cdot 10^{-4}$
POLR1A	16	$6.90 \cdot 10^{-5}$	$1.35 \cdot 10^{-2}$	$3.97 \cdot 10^{-4}$	$3.10 \cdot 10^{-4}$
CNGA3	26	$2.37 \cdot 10^{-4}$	$1.14 \cdot 10^{-3}$	$1.09 \cdot 10^{-4}$	$1.12 \cdot 10^{-3}$
XPOT	9	$5.53 \cdot 10^{-4}$	$1.76 \cdot 10^{-2}$	$1.60 \cdot 10^{-4}$	$9.93 \cdot 10^{-4}$
VWA3B	51	$6.06 \cdot 10^{-4}$	$9.39 \cdot 10^{-2}$	$2.06 \cdot 10^{-4}$	$1.86 \cdot 10^{-3}$
C11orf49	24	$2.39 \cdot 10^{-4}$	$3.49 \cdot 10^{-3}$	$3.84 \cdot 10^{-4}$	$3.41 \cdot 10^{-3}$
MMRN1	10	$4.54 \cdot 10^{-4}$	$9.35 \cdot 10^{-3}$	$3.83 \cdot 10^{-2}$	$2.86 \cdot 10^{-4}$
DGKQ	9	$3.98 \cdot 10^{-4}$	$6.45 \cdot 10^{-3}$	$7.41 \cdot 10^{-3}$	$2.95 \cdot 10^{-4}$
SCARB2	22	$5.62 \cdot 10^{-4}$	$6.80 \cdot 10^{-2}$	$7.08 \cdot 10^{-4}$	$4.19 \cdot 10^{-4}$
TMEM175	10	$5.76 \cdot 10^{-4}$	$1.11 \cdot 10^{-2}$	$3.52 \cdot 10^{-3}$	$4.22 \cdot 10^{-4}$
HCN1	36	$8.65 \cdot 10^{-4}$	$8.14 \cdot 10^{-3}$	$1.85 \cdot 10^{-2}$	$4.24 \cdot 10^{-4}$
AGMAT	5	$4.83 \cdot 10^{-4}$	$3.26 \cdot 10^{-3}$	$4.62 \cdot 10^{-4}$	$5.62 \cdot 10^{-4}$
NTSR1	32	$4.74 \cdot 10^{-4}$	$7.17 \cdot 10^{-3}$	$7.13 \cdot 10^{-3}$	$2.43 \cdot 10^{-3}$

Table 3.2: P-values from the CGEM breast cancer GWAS for all four methods are reported. The list is sorted in increasing order based on the smallest of the four P-values.

associations.

The Q-Q plot based on the GHC gene-level P-values for this GWAS are presented in Figure 3.5. For the most significant gene, FGFR2, the GHC had the smallest P-value of $2.41 \cdot 10^{-5}$. This P-value should not be directly compared with SNP-level P-values because there is less of a multiple testing problem (528173 SNPs compared to 14991 genes). It should also be noted that the iHC P-value for FGFR2 is 0.107 which reflects the attenuated marginal test statistics seen in Figure 3.1. The second most significant gene, TBK1, is closely related to IKBKE, a known breast cancer oncogene (Boehm et al., 2007). For TBK1, SKAT detected the association with the smallest P-value of $5.63 \cdot 10^{-5}$. The third most significant gene, PTCD3, has previously been identified in a gene network significantly associated with breast cancer (Jia et al., 2011). For PTCD3, the GHC had the most significant P-value of $5.77 \cdot 10^{-5}$.

Overall for the most significant genes (ordered by the minimum P-value of all four methods), MinP and iHC tended to yield less significant P-values than GHC and SKAT. SKAT and GHC both showed similar strength in detecting these top associations, but

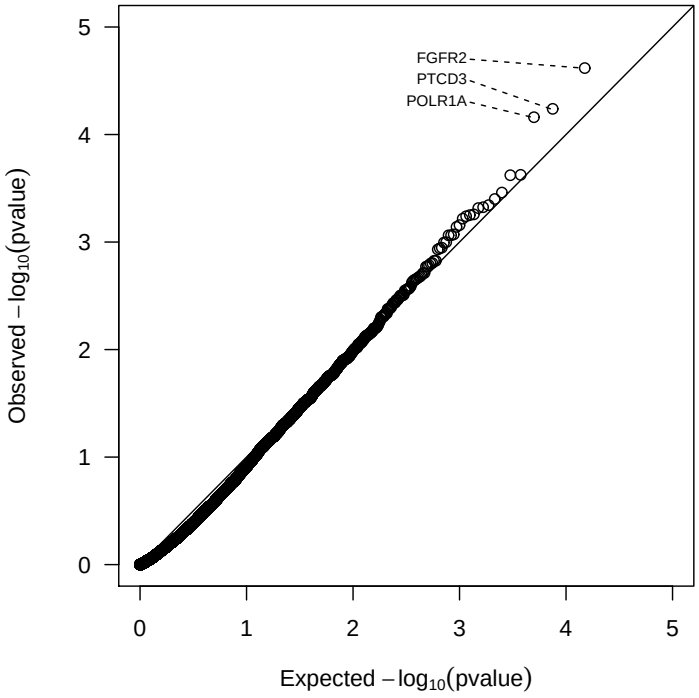

Figure 3.5: Q-Q plot of GHC P-values for the SNP-set testing of the CGEM breast cancer GWAS data. SNP-sets were constructed at the gene-level, also including SNPs within 20kb from the border of each gene. SNP-sets with 4 or fewer SNPs were not included in the analysis leading to total of 14991 SNP-sets evaluated.

SKAT had some very insignificant P-values greater than 0.01 for a few of the top 15 genes (MMRN1 and HCN1) whereas GHC never had a P-value greater than $8.7 \cdot 10^{-4}$. This difference demonstrates how GHC is more robust than SKAT to the LD in the top genes and is a safer bet in genome-wide association studies where there is a great variety of LD structures encountered.

3.8 Discussion

In this paper, the higher criticism, a popular sparse signal detection method originally designed for high dimensional problems, is generalized to the setting of SNP-set testing in genetic association studies. Unlike the original HC, our GHC is flexible to SNP-sets of arbitrary size and LD structure. We propose an analytic method to compute the p-values of GHC for finite samples that is computationally efficient and requires nei-

ther simulation nor asymptotics in p to obtain its P-values. This is advantageous when scanning a large number of genes in GWAS. An implementation of the method is freely available for use in the *R* package, *GHC*. We showed through simulation and analysis of the CGEM breast cancer GWAS data that the GHC is more robust to varied LD structures than competing methods. In particular, we show that the GHC is more powerful than the iHC regardless of the LD structure, and as a result should always be the preferred method.

As demonstrated by the simulation studies and the analysis of the breast cancer GWAS data, both GHC and SKAT complement each other well. GHC tends to outperform SKAT in the high sparsity settings while SKAT outperforms GHC when sparsity is low. This suggests that an omnibus test that combines the strengths of both GHC and SKAT could potentially be a powerful alternative in a variety of scenarios.

Though the GHC does not require asymptotics in p, asymptotics in the sample size N *is* assumed. The marginal test statistics are assumed to be normally distributed, but this is a poor assumption if SNPs are rare or if the sample size is small. For GWAS this is not a problem due to their tendency to have large cohorts and only common SNPs genotyped. However, if GHC is to be extended to sequencing studies where SNPs can be rare, then the normality assumption of the marginal test statistics must be relaxed. Advances in high-throughput sequencing technology are reshaping the field of genetics research, with more sequencing studies such as the 1000 Genomes Project and the NHGRI Genome Sequencing Program. It is of future research interest to extend the GHC to sequencing studies by using a different marginal test statistic that is robust to rare SNPs.

References

AHDESMÄKI, M., STRIMMER, K. ET AL. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics*, **4** 503–519.

ARIAS-CASTRO, E., CANDÈS, E. and PLAN, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, **39** 2533–2556.

BANSAL, V., LIBIGER, O., TORKAMANI, A. and SCHORK, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, **11** 773–785.

BARNETT, I. and LIN, X. (2013). Analytic p-value calculation for the higher criticism test in finite *p* problems. *Manuscript in preparation*.

BASU, S. and PAN, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic epidemiology*, **35** 606–619.

BIESECKER, L. G., SHIANNA, K. V. and MULLIKIN, J. C. (2011). Exome sequencing: the expert view. *Genome Biol*, **12** 128.

BOEHM, J. S., ZHAO, J. J., YAO, J., KIM, S. Y., FIRESTEIN, R., DUNN, I. F., SJOSTROM, S. K., GARRAWAY, L. A., WEREMOWICZ, S., RICHARDSON, A. L. ET AL. (2007). Integrative genomic approaches identify *IKBKE* as a breast cancer oncogene. *Cell*, **129** 1065–1079.

CHEN, Z., ZHENG, G., GHOSH, K. and LI, Z. (2005). Linkage disequilibrium mapping of

quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, **77** 661–669.

CHEVERUD, J. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, **87** 52–58.

CIRULLI, E. T. and GOLDSTEIN, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, **11** 415–425.

CLÉMENT, K., VAISSE, C., MANNING, B. S. J., BASDEVANT, A., GUY-GRAND, B., RUIZ, J., SILVER, K. D., SHULDINER, A. R., FROGUEL, P. and STROSBERG, A. D. (1995). Genetic variation in the β3-adrenergic receptor and an increased capacity to gain weight in patients with morbid obesity. *New England Journal of Medicine*, **333** 352–354.

CONNEELY, K. and BOEHNKE, M. (2007). So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics*, **81** 1158–1168.

DAVIES, R. (1980). Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, **29** 323–333.

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, **32** 962–994.

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, **105** 14790–14795.

GAENSSLER, P. (1983). *Empirical processes*. Institute of Mathematical Statistics.

GU, C., TODOROV, A. and RAO, D. (1997). Genome screening using extremely discordant and extremely concordant sib pairs. *Genetic epidemiology*, **14** 791–796.

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, **38** 1686–1732.

HAN, B., KANG, H. and ESKIN, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, **5** e1000456.

HERNANDEZ, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24** 2786–2787.

HOEFFDING, W. and ROBBINS, H. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal*, **15** 773–780.

HU, S., ZHONG, Y., HAO, Y., LUO, M., ZHOU, Y., GUO, H., LIAO, W., WAN, D., WEI, H., GAO, Y. ET AL. (2009). Novel rare alleles of abca1 are exclusively associated with extreme high-density lipoprotein-cholesterol levels among the han chinese. *Clinical Chemistry and Laboratory Medicine*, **47** 1239–1245.

HUANG, B. and LIN, D. (2007). Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, **80** 567–576.

HUNTER, D., KRAFT, P., JACOBS, K., COX, D., YEAGER, M., HANKINSON, S., WACHOLDER, S., WANG, Z., WELCH, R., HUTCHINSON, A. ET AL. (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, **39** 870–874.

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics*, **4** 1476–1526.

IOANNIDIS, J. P., THOMAS, G. and DALY, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, **10** 318–329.

JAESCHKE, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *The Annals of Statistics*, **7** 108–115.

JI, W., FOO, J. N., O'ROAK, B. J., ZHAO, H., LARSON, M. G., SIMON, D. B., NEWTON-CHEH, C., STATE, M. W., LEVY, D., LIFTON, R. P. ET AL. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature genetics*, **40** 592–599.

JIA, P., ZHENG, S., LONG, J., ZHENG, W. and ZHAO, Z. (2011). dmgwas: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, **27** 95–102.

KHOR, C. C. and GOH, D. L.-M. (2010). Strategies for identifying the genetic basis of dyslipidemia: genome-wide association studies vs. the resequencing of extremes. *Current opinion in lipidology*, **21** 123–127.

KIMMEL, G. and SHAMIR, R. (2006). A fast method for computing high-significance disease association in large population-based studies. *The American Journal of Human Genetics*, **79** 481–492.

LEE, S., EMONDS, M., BAMSHAD, M., BARNES, K., RIEDER, M., NICKERSON, D., CHRISTIANI, D., WURFEL, M. and LIN, X. (2012a). Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*.

LEE, S., WU, M. and LIN, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*.

LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, **83** 311–321.

LI, D., LEWINGER, J. P., GAUDERMAN, W. J., MURCRAY, C. E. and CONTI, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic epidemiology*, **35** 790–799.

LIANG, K.-Y., HUANG, C.-Y. and BEATY, T. H. (2000). A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. *The American Journal of Human Genetics*, **66** 1631–1641.

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84** 309–326.

LIU, D., GHOSH, D. and LIN, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics*, **9** 292.

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, **63** 1079–1088.

MACCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models*, vol. 37. CRC press.

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, **5** e1000384.

MAHER, B. (2008). The case of the missing heritability. *Nature*, **456** 18–21.

MANOLIO, T., COLLINS, F., COX, N., GOLDSTEIN, D., HINDORFF, L., HUNTER, D., MCCARTHY, M., RAMOS, E., CARDON, L., CHAKRAVARTI, A. ET AL. (2009). Finding the missing heritability of complex diseases. *Nature*, **461** 747–753.

MORGENTHALER, S. and THILLY, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **615** 28–56.

MORRIS, A. P. and ZEGGINI, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, **34** 188–193.

MOSKVINA, V. and SCHMIDT, K. (2008). On multiple-testing correction in genome-wide association studies. *Genetic epidemiology*, **32** 567–573.

MUKHERJEE, R., PILLAI, N. S. and LIN, X. (2013). Hypothesis testing for sparse binary regression. *arXiv preprint arXiv:1308.0764*.

NEALE, B. M., RIVAS, M. A., VOIGHT, B. F., ALTSHULER, D., DEVLIN, B., ORHO-MELANDER, M., KATHIRESAN, S., PURCELL, S. M., ROEDER, K. and DALY, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, **7** e1001322.

NEJENTSEV, S., WALKER, N., RICHES, D., EGHOLM, M. and TODD, J. A. (2009). Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324** 387–389.

NG, P. C., LEVY, S., HUANG, J., STOCKWELL, T. B., WALENZ, B. P., LI, K., AXELROD, N., BUSAM, D. A., STRAUSBERG, R. L. and VENTER, J. C. (2008). Genetic variation in an individual human exome. *PLoS genetics*, **4** e1000160.

NYHOLT, D. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, **74** 765–769.

PAN, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic epidemiology*, **33** 497–507.

PRICE, A. L., KRYUKOV, G. V., DE BAKKER, P. I., PURCELL, S. M., STAPLES, J., WEI, L.-J. and SUNYAEV, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, **86** 832–838.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38** 904–909.

PRICE, R. A., LI, W.-D. and ZHAO, H. (2008). Fto gene snps associated with extreme obesity in cases, controls and extremely discordant sister pairs. *BMC medical genetics*, **9** 4.

RAMSER, J., AHEARN, M. E., LENSKI, C., YARIZ, K. O., HELLEBRAND, H., VON RHEIN, M., CLARK, R. D., SCHMUTZLER, R. K., LICHTNER, P., HOFFMAN, E. P. ET AL. (2008). Rare missense and synonymous variants in ube1 are associated with x-linked infantile spinal muscular atrophy. *The American Journal of Human Genetics*, **82** 188–193.

RISCH, N. and ZHANG, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, **268** 1584–1589.

SCHWARTZMAN, A. and LIN, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, **98** 199–214.

SLATKIN, M. (1999). Disequilibrium mapping of a quantitative-trait locus in an expanding population. *The American Journal of Human Genetics*, **64** 1765–1773.

SPITZ, M. R., AMOS, C. I., DONG, Q., LIN, J. and WU, X. (2008). The chrna5-a3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *Journal of the National Cancer Institute*, **100** 1552–1556.

TZENG, J.-Y. and ZHANG, D. (2007). Haplotype-based association analysis via variance-components score test. *The American Journal of Human Genetics*, **81** 927–938.

VICTOR, R. G., HALEY, R. W., WILLETT, D. L., PESHOCK, R. M., VAETH, P. C., LEONARD, D., BASIT, M., COOPER, R. S., IANNACCHIONE, V. G., VISSCHER, W. A. ET AL. (2004). The dallas heart study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *The American journal of cardiology*, **93** 1473–1480.

WALLACE, C., CHAPMAN, J. M. and CLAYTON, D. G. (2006). Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, **78** 498–504.

WANG, C.-I., CHIEN, K.-Y., WANG, C.-L., LIU, H.-P., CHENG, C.-C., CHANG, Y.-S., YU, J.-S. and YU, C.-J. (2012). Quantitative proteomics reveals regulation of kpna2 and its potential novel cargo proteins in non-small cell lung cancer. *Molecular & Cellular Proteomics* mcp–M111.

WU, M., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011a). Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT). *Manuscript*.

WU, M., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011b). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, **89** 82–93.

WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. and LIN, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, **86** 929–942.

ZHANG, Y. and LIU, J. (2011). Fast and accurate approximation to significance tests in genome-wide association studies. *Journal of the American Statistical Association*, **106** 846–857.

ZIENOLDDINY, S., SKAUG, V., LANDVIK, N. E., RYBERG, D., PHILLIPS, D. H., HOULSTON, R. and HAUGEN, A. (2009). The tert-clptm1l lung cancer susceptibility variant associates with higher dna adduct formation in the lung. *Carcinogenesis*, **30** 1368–1371.

Appendix A

Power and null distribution derivations for CEP-SKAT-O

A.1 Rare causal variants are enriched in phenotypic extremes

We consider a phenotype of the ith individual to be modeled as

$$y_i = \alpha_0 + \boldsymbol{X}'\boldsymbol{\alpha} + \boldsymbol{G}_i'\boldsymbol{\beta} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Here α_0 is an intercept term with $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_m]'$ as the vector of regression coefficients for the covariates $\mathbf{X}_i$, and $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]'$ as the vector of regression coefficients for the p genetic variants $\boldsymbol{G}_i = (G_{i1}, \cdots, G_{ip})'$. First we will calculate the MAF when sampling extremes where there is a single causal variant and no covariate, i.e., $p = 1$ and $m = 0$. We will then extend the results to multiple causal variants and derive an analytic relationship between the MAF of a causal variant using extreme phenotype sampling to the background population MAF.

Under the single causal variant/no covariate model $y = \beta G + \epsilon$, we desire to show that an individual's probability of having at least one minor allele of the causal variant is increased when their phenotype is extreme. We assume the additive model, but the results can be easily extended to the dominant-recessive model. Without loss of generality, we consider the case where the causal variant has a positive effect on phenotype ($\beta > 0$), and show that for $c > 0$, $Pr(G > 0|y > c) > Pr(G > 0)$.

We first write

$$Pr(G > 0|y > c) = Pr(G > 0)\frac{Pr(y > c|G > 0)}{Pr(y > c)}.$$

Hence showing $Pr(G > 0|y > c) > Pr(G > 0)$ is equivalent to showing $Pr(y > c|G > 0) > Pr(y > c)$. We condition on G to achieve the desired result:

$$\begin{aligned} Pr(y > c|G > 0) - Pr(y > c) &= Pr(y > c|G > 0) - Pr(y > c|G = 0)Pr(G = 0) \\ &\quad -Pr(y > c|G > 0)Pr(G > 0) \\ &= \{Pr(y > c|G > 0) - Pr(y > c|G = 0)\}Pr(G = 0) \\ &> \{Pr(y > c|G = 1) - Pr(y > c|G = 0)\}Pr(G = 0) \\ &= \{Pr(\beta + \epsilon > c) - Pr(\epsilon > c)\}Pr(G = 0) > 0 \end{aligned}$$

Hence the MAF of a causal variant among extreme phenotypes is higher than that in the population. One can easily see that if G is not a causal variant, i.e., $\beta = 0$, then $P(G > 0|y > c) = P(G > 0)$, i.e., the MAF by sampling extremes is the same as in the population.

We next calculate the expected MAF of a causal rare variant in the presence of single or multiple causal variants under extreme phenotype sampling. Specifically, we show below that the expected MAF of a causal variant in extreme phenotype samples can be written as a function of the MAFs of the p causal variants in the background population, the threshold and the regression coefficients β's. Consider the no-covariate model

$$y = \beta_1 G_1 + \cdots + \beta_p G_p + \epsilon. \tag{A.1}$$

We are interested in estimating $Pr(G_j = g|y > c)$ for $g = 0$, 1, and 2. For simplicity, we assume in our analytic calculations no LD between causal variants which is a plausible assumption when variants are rare. This assumption gives us:

$$Pr(G_1 = g_1, ..., G_p = g_p) = \prod_{l=1}^{p} Pr(G_l = g_l).$$

Note that each $Pr(G_l = g_l)$ can be easily estimated from the data. We model the effect

size of the jth causal variant as $\beta_j = -a \cdot log_{10}(MAF_j)$ for some constant $a > 0$, i.e., we assume positive effects.

Write

$$Pr(G_j = g|y > c) = Pr(y > c|G_j = g)\frac{Pr(G_j = g)}{Pr(y > c)}, \tag{A.2}$$

which means we need just compute $P(y > c)$ and $P(y > c|G_j = g)$. This can be done by conditioning on the remaining causal variants as

$$\begin{aligned} Pr(y > c) &= Pr(\sum_{l=1}^{p} \beta_l G_l + \epsilon > c) \\ &= \sum_{i_1=0}^{2} \cdots \sum_{i_p=0}^{2} Pr\left\{\sum_{l=1}^{p} \beta_l G_l + \epsilon > c | G_1 = g_1, ..., G_p = g_p\right\} Pr(G_1 = g_1, ..., G_p = g_p) \\ &= \sum_{i_1=0}^{2} \cdots \sum_{g_p=0}^{2} Pr(\sum_{l=1}^{p} \beta_l g_l + \epsilon > c) \prod_{l=1}^{p} P(G_l = g_l) \\ &= \sum_{g_1=0}^{2} \cdots \sum_{g_p=0}^{2} \Phi(\sum_{l=1}^{p} \beta_l g_l - c) \prod_{l=1}^{p} Pr(G_l = g_l) \end{aligned} \tag{A.3}$$

Calculations for $P(y > c|G_j = g)$ are identical except for there being no need to condition on the jth variant:

$$Pr(y > c|G_j = g) = \sum_{g_1=0}^{2} \cdots \sum_{g_{j-1}=0}^{2} \sum_{g_{j+1}=0}^{2} \cdots \sum_{g_p=0}^{2} \Phi(a\beta_j + \sum_{l \neq j} \beta_l g_l - c) \prod_{l \neq j} P(G_l = g_l) \tag{A.4}$$

It follows from (A.2) that we can calculate the expected MAF in extreme phenotype samples as a function of the MAF of the causal variants, the threshold c, and the regression coefficients β_j's in the phenotype model (A.1) as

$$Pr(G_j > 1|y > c) = E(G_j|y > c)/2 = 0 * Pr(G_j = 0|y > c) + 0.5 * Pr(G_j = 1|y > c) + 1 * Pr(G_j = 2|y > c).$$

One can also use equations (A.2) and (A.4) to easily show that $Pr(G_j = g|y > c) > Pr(G_j = g)$ if β's are not equal to 0, i.e., the MAF of a causal variant is higher in extreme phenotype samples than their population counterpart.

A.2 Null distribution of Continuous Extreme Phenotype SKAT

Suppose the true NULL model is

$$y_i = \boldsymbol{X}_i'\boldsymbol{\alpha} + \epsilon \tag{A.5}$$

where $\boldsymbol{X}_i = (x_{i0}, x_{i1}, \ldots, x_{im})$ is the covariates of i^{th} individual with $x_{i0} = 1$, and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_m)'$ is a vector of regression coefficients of $\boldsymbol{X}_i$, and $\epsilon \sim N(0, \sigma^2)$. Note that we use a slightly different notation in which $\boldsymbol{X}_i$ includes the intercept. Suppose we select n samples with either $y_i > c_1$ or $y_i < c_2$, and denote the selected y_i as y_i^* . For notational simplicity, we use i to indicate the selected individuals. Under the null hypothesis, y_i^* follows a truncated Gaussian distribution with the density function

$$f(y_i^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{exp\{-(y_i^* - \boldsymbol{X}_i'\boldsymbol{\alpha})^2/2\sigma^2\}}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})}, \tag{A.6}$$

where $t_{i1} = (c_1 - \boldsymbol{X}_i'\boldsymbol{\alpha})/\sigma$ and $t_{i2} = (c_2 - \boldsymbol{X}_i'\boldsymbol{\alpha}))/\sigma$. The first derivative of log likelihood function is

$$u_j = \frac{\partial \ell}{\partial \alpha_j} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{X}_i'\boldsymbol{\alpha} + m_i),$$

and the second derivative is

$$J_{ik} = \frac{\partial^2 \ell}{\partial \alpha_j \partial \alpha_k} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}x_{ik}(-1 + v_i),$$

where

$$m_i = \sigma \frac{\phi(t_{i2}) - \phi(t_{i1})}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})}, \quad \text{and} \quad v_i = \frac{t_{i2}\phi(t_{i2}) - t_{i1}\phi(t_{i1})}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})} + \frac{m_i^2}{\sigma^2}.$$

Define $\boldsymbol{S} = -\boldsymbol{J}$, $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)'$, $\boldsymbol{u} = (u_0, \ldots, u_m)'$, and $\boldsymbol{m} = (m_1, \ldots, m_n)'$. By the Fisher Scoring (or Newton Raphson) procedure, new $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha}^* = \boldsymbol{\alpha} + \boldsymbol{S}^{-1}\boldsymbol{u},$$

Type I error estimates for Continuous Extreme Phenotype (CEP-SKAT-O)

α-level	0.05	0.01	1×10^{-5}	2.5×10^{-6}	1×10^{-6}
n=500	0.0471	0.0097	1.66×10^{-5}	4.22×10^{-6}	1.70×10^{-6}
n=1000	0.0480	0.0100	1.48×10^{-5}	4.70×10^{-6}	2.19×10^{-6}
n=2000	0.0497	0.0104	1.70×10^{-5}	4.75×10^{-6}	2.21×10^{-6}

Table A.1: Phenotypes were simulated under the null model (A.5) using two covariates and added Gaussian noise, but with no genotype effects. Estimates are based on 20 million simulated p-values. Adjusted SKAT-O was used to adjust the p-value for small sample size.

hence $\boldsymbol{S}(\alpha^* - \alpha) = \boldsymbol{u}$. Since $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}/\sigma^2$, where $\boldsymbol{V} = diag\{(1 - v_i)\}$ and $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots \boldsymbol{X}_n]'$,

$$\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) = \boldsymbol{X}'(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{m}).$$

Now we can treat the Fisher scoring algorithm as a weighted least square problem. Define a working vector

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{V}^{-1}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{m}),$$

and then $\boldsymbol{\alpha}^*$ is a weighted least square estimtor of the linear model $\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}}$ with $E(\tilde{\boldsymbol{\epsilon}}) = 0$ and $Var(\tilde{\boldsymbol{\epsilon}}) = \boldsymbol{V}^{-1}$. Since $E(y_i^*) = \boldsymbol{X}_i'\boldsymbol{\alpha} - m_i$, the SKAT test statistic with linear weighted kernel is

$$\begin{aligned} Q_S &= (\boldsymbol{y}^* - \hat{\mu})'\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'(\boldsymbol{y}^* - \hat{\mu}) = (\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\alpha}^*)'\boldsymbol{V}\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'\boldsymbol{V}(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\alpha}^*) \\ &= \tilde{\boldsymbol{Y}}\boldsymbol{P}_0\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'\boldsymbol{P}_0\tilde{\boldsymbol{Y}}, \end{aligned}$$

where $\boldsymbol{P}_0 = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}$. From $Var(\tilde{y}_i) = (1 - v_i)\sigma^2$, the asymptotic null distribution of Q_S is

$$\sum \lambda_v \chi_v^2,$$

where λ_v is the v^{th} eigenvalue of $\hat{\sigma}^2\boldsymbol{P}_0^{1/2}\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'\boldsymbol{P}_0^{1/2}$.

Calculations of Q_S require fitting the null model (A.5) using extreme phenotypes y_i^* under truncated normal likelihood (A.6). The Newton-Raphson method can be used to estimate $\boldsymbol{\alpha}$ and σ^2.

A.3 Null distribution of the optimal unified test for continuous extreme phenotype

Suppose

$$Q_\rho = (1-\rho)Q_S + \rho Q_B,$$

which is the test statistic of the proposed unified test, where Q_S is the SKAT statistic and Q_B is the burden test statistic. The class of test statistics Q_ρ includes both SKAT ($\rho = 0$) and the burden test ($\rho = 1$) as special cases, and p_ρ is a p-value computed based on Q_ρ. Then, the test statistic of the optimal test is

$$T = min\{p_{\rho_1}, \ldots, p_{\rho_b}\}, \quad 0 = \rho_1 < \rho_2 < \ldots < \rho_b = 1. \tag{A.7}$$

Define $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}\boldsymbol{G}\boldsymbol{W}$, $\bar{\boldsymbol{z}} = (\bar{z}_1, \ldots, \bar{z}_n)'$, where $\bar{\boldsymbol{z}}_i = \sum_{j=1}^{p} z_{ij}/p$, and $\boldsymbol{M} = \bar{\boldsymbol{z}}(\bar{\boldsymbol{z}}'\bar{\boldsymbol{z}})^{-1}\bar{\boldsymbol{z}}'$. We further let

$$\tau(\rho) = p^2\rho\bar{\boldsymbol{z}}'\bar{\boldsymbol{z}} + \frac{1-\rho}{\bar{\boldsymbol{z}}'\bar{\boldsymbol{z}}}\sum_{j=1}^{p}(\bar{\boldsymbol{z}}'\boldsymbol{z}_{.j})^2,$$

where $\boldsymbol{z}_{.j}$ is the j^{th} column of $\boldsymbol{Z}$. Following the same argument in Lee *et al.*(2012), it can be shown that Q_ρ is asymptotically equivalent as

$$(1-\rho)(\sum_{k=1}^{m}\tilde{\lambda}_k\eta_k + \zeta) + \tau(\rho)\eta_0, \tag{A.8}$$

where $\{\tilde{\lambda}_1, \ldots \tilde{\lambda}_m\}$ are non-zero eigenvalues of $\boldsymbol{Z}'(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Z}$, $\eta_k (k = 0, \ldots, m)$ are i.i.d χ^2_1 random variables, and ζ satisfies the following conditions:

$$E(\zeta) = 0, \quad Var(\zeta) = 4trace(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}'(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Z}),$$
$$Corr(\sum_{k=1}^{m}\lambda_k\eta_k, \zeta) = 0, \quad \text{and} \quad Corr(\eta_0, \zeta) = 0.$$

It shows that the Q_ρs are mixtures of shared random variables, and the only differences among different Q_ρs are the mixing coefficients. From this fact, a p-value of T can be efficiently computed through one dimensional numerical integration. Details can be found in Lee *et al.*(2012) (Lee et al., 2012b).

A.4 Small Sample Adjustment

It is known that the SKAT family tests can produce conservative results when the trait is binary and the sample size is small. The same conservative pattern can be observed when we test extreme continuous phenotypes. To resolve this issue, we adopt the same strategy as that in Lee *et al.*(2012) (Lee et al., 2012a) in which we adjust asymptotic null distribution of the test statistics by estimating small sample variance and kurtosis. To estimate these moments, we generate resampled test statistics using the parametric bootstrap approach. In particular, B sets of truncated normal random variables are generated from the model (A.6) with estimated $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}$ under the null hypothesis, and the variance and kurtosis of Q_S and Q_B are estimated using resampled phenotype sets. Then, we apply the same algorithm in Lee *et al.*(2012) (Lee et al., 2012a).

A.5 Theoretical Power Calculation

Power calculation derivations are available for SKAT and SKAT-O, but adjustments need to be made to account for extreme phenotype sampling. Derivations for power calculations for continuous extreme phenotype SKAT (CEP-SKAT-O) mirror Lee *et al.*(2012) Lee et al. (2012b) in their calculation for continuous phenotypes, but a key distinction in the treatment of the genotype matrix needs to be made. By sampling from the extremes, causal variants tend to occur more often in the sample than observed in the population, and the power calculation should reflect this bias appropriately. In a random sample the MAF of all sampled variants should be consistent for their respective population MAFs (note that rate of convergence is slow for rare variants). In an EPS sample, consistency is not achieved and the likelihood of sampling a genotype must be adjusted accordingly in the power calculation.

To account for this biased sampling of genotypes, a reduced genotype matrix $\boldsymbol{G}_R$ is used in the calculation of the SKAT statistic. We define $\boldsymbol{G}_R = \boldsymbol{B}\boldsymbol{G}$ where $\boldsymbol{B}$ is an n by n diagonal matrix with jth diagonal equal to $\sqrt{P(\text{Variant j exists in the EPS sample})}$. We can calculate $\boldsymbol{B}$ given the upper and lower cutoffs for sampling extreme phenotypes by

taking the tail probabilities beyond these cutoffs of the normal distribution with a mean of the jth entry of $\boldsymbol{G}\beta$ and a variance of 1. The resulting test statistic is:

$$Q_S = (\boldsymbol{y}^* - \hat{\mu}_R)' \boldsymbol{G}_R \boldsymbol{W}_\rho \boldsymbol{G}_R' (\boldsymbol{y}^* - \hat{\mu}_R)$$

where $\hat{\mu}_R$ is the expected value of the truncated normal $\boldsymbol{y}^*$ and is a function of $\boldsymbol{G}_R$. $\boldsymbol{W}_\rho$ is the matrix of weights adjusted by ρ through the matrix $\boldsymbol{R}_\rho$ as done in Lee *et al.*(2012) Lee et al. (2012b) who also recommend to approximate ρ using the percent of causal variants and percent of variants with a positive effect on phenotype. The distribution of Q_S under the alternative hypothesis can then be approximated by a non-central χ^2 distribution. The distribution of Q_S under the null hypothesis is approximated in a similar manner except under the assumption of no variant effects. Power is estimated by the area in the upper tail of the alternative distribution that lies above the critical value taken from the null distribution.

Power estimates are obtained by averaging the estimated power over many randomly selected regions of equivalent size (we selected 3kb regions) in order to account for variability of genotypes by region. In each region a new $\boldsymbol{G}_R$ is chosen with individuals selected based on their probability of being observed in the phenotypic extremes of the sample given their genotypes. We compare the theoretical powers and empirical powers in Figure 1.4 and Figure A.1.

A.6 DHS data analysis sensitivity to different cutoffs

In the Dallas Heart Study, we examine how the p-values for all the EPS methods are affected by altering the extreme phenotype cutoff, and results are presented in Figure A.2. DEP-N represents DEP-Burden test in the main manuscript. Note that two additional burden tests, DEP-W and DEP-C are included. DEP-W uses a weighted count with beta(1,25) weights while DEP-C is an adaptation of CAST for EPS. We range the cutoffs from 15% to 30% in 1% increments to capture the sensitivity of the tests to different cutoffs. From the spikes in each methods p-values over slight changes in this cutoff value, it is clear that inclusion or exclusion of certain individuals could affect the overall signif-

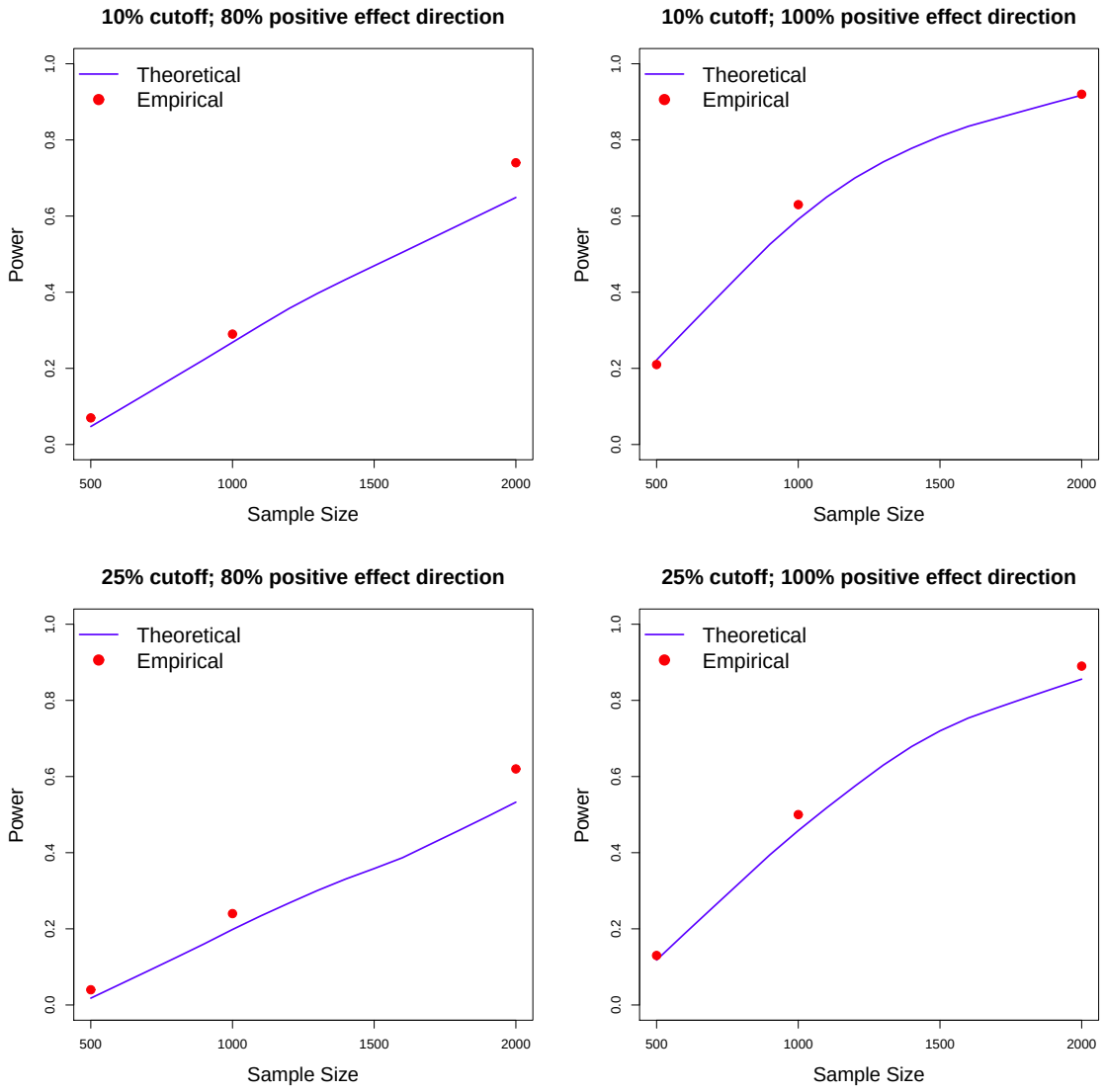

Figure A.1: In this setting, 60% of variants were considered causal in a 3kb region. Theoretical power for optimal continuous extreme phenotype SKAT (CEP-SKAT-O) is compared with the empirical power estimated using 300 simulations for each estimate. Four settings are considered: sampling 10% and 20% high/low extreme phenotypes; 80%/20% causal variants have positive/negative effects and 100% causal variants have positive effects.

icance. It is important to note that because we have a fixed sample size of 3476, smaller cutoffs lead a smaller sizes through EPS. As an example, a tail cutoff of 15% leads to half the EPS sample size that we would see with a tail cutoff of 30%. Because significance is sensitive to sample size, direct comparison between p-values at different cutoffs is not

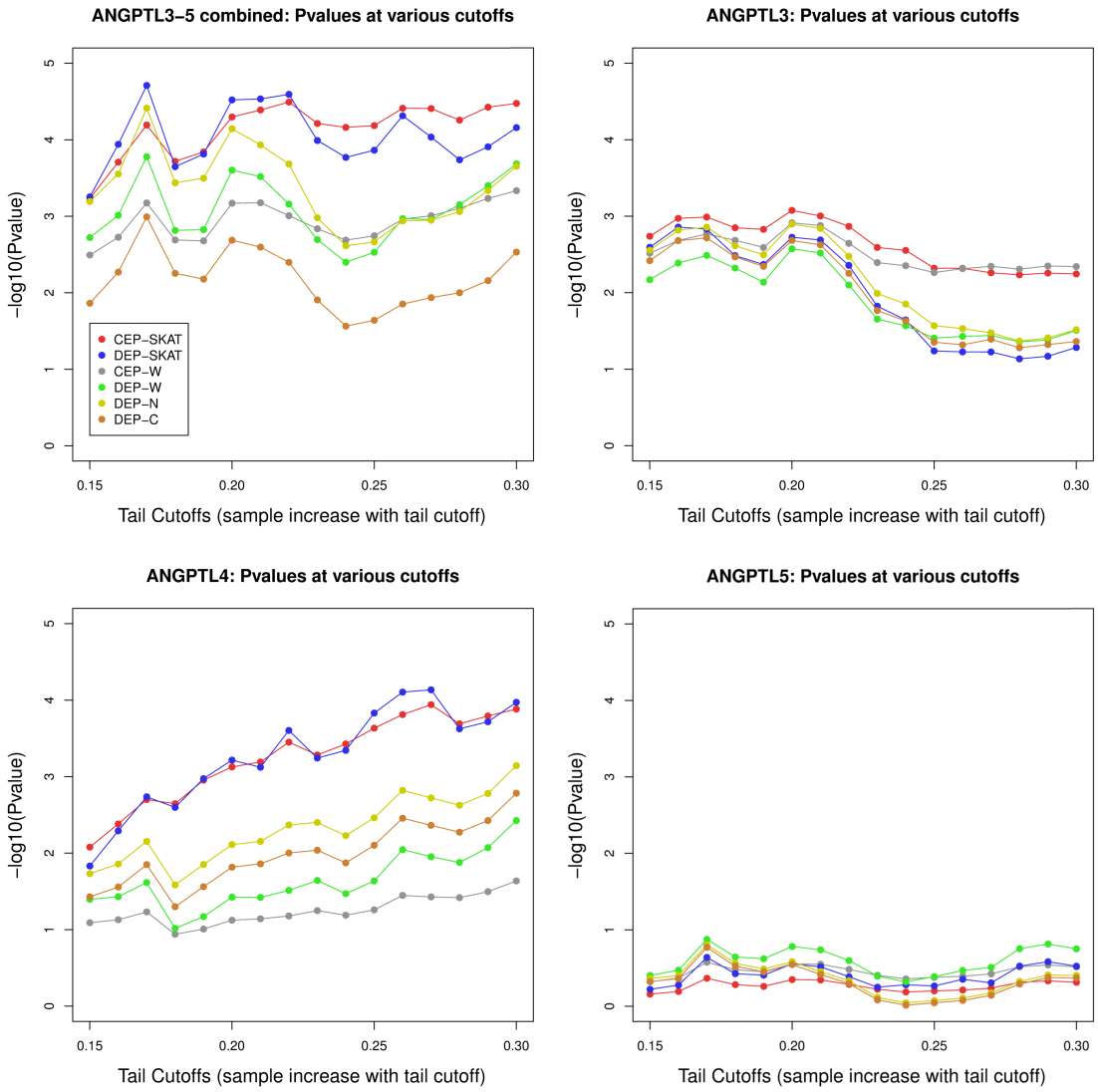

Figure A.2: The p-values using six EPS association tests using different extreme cutoffs are demonstrated. Each of the three genes, ANGPTL3, ANGPTL4, and ANGPTL5 are tested separately. A test combining all three genes is also included.

appropriate in Figure A.2. This analysis is presented for each of ANGPTL3, ANGPTL4, and ANGPTL5 being tested for separately as well as a combined analysis across the three genes.

For the three-gene combined analysis, CEP-SKAT-O gives the smallest p-value than the other methods when the cutoff of selecting extreme phenotypes is less than 22% and slightly higher p-values than DEP-SKAT-O when the cutoff is greater than 22%. For

Analysis of the Dallas Heart Study triglyceride data

n=1389	CEP-SKAT-O	DEP-SKAT-O	RS-SKAT-O	DEP-W	DEP-N	DEP-C
20%	5.0×10^{-5}	3.0×10^{-5}	1.3×10^{-2}	2.5×10^{-4}	7.2×10^{-5}	2.1×10^{-3}
30%	1.0×10^{-3}	1.9×10^{-3}	1.3×10^{-2}	2.0×10^{-3}	2.8×10^{-3}	1.7×10^{-2}
40%	8.9×10^{-3}	1.2×10^{-2}	1.3×10^{-2}	1.2×10^{-2}	1.9×10^{-2}	4.5×10^{-2}

Table A.2: Analysis results of the Dallas Heart Study (DHS) sequence data using various test methods and sampling schemes. The DHS sequenced 3,476 subjects. A total 1,389 individuals were selected with highest and lowest 20% logTG levels in each age-gender spectrum. For sampling with higher cutoffs (30% and 40%), 1389 individuals were randomly sub-sampled among the individuals belongs to larger tails. In these cases, median p-values are presented in the table from 1000 sampling iterations. Since the sample size was large, the small sample adjustment was not applied. DEP-N represents DEP-Burden test in the main manuscript. Two additional burden tests are included: DEP-W uses a weighted count while DEP-C is an adaptation of CAST for EPS.

ANGPL3 gene, CEP-SKAT-O overall has the smallest p-compared with the other methods. For ANGPTL4 and ANGPTL5 genes, CEP-SKAT-O has similar p-values to DEP-SKAT-O. Both outperform the burden tests.

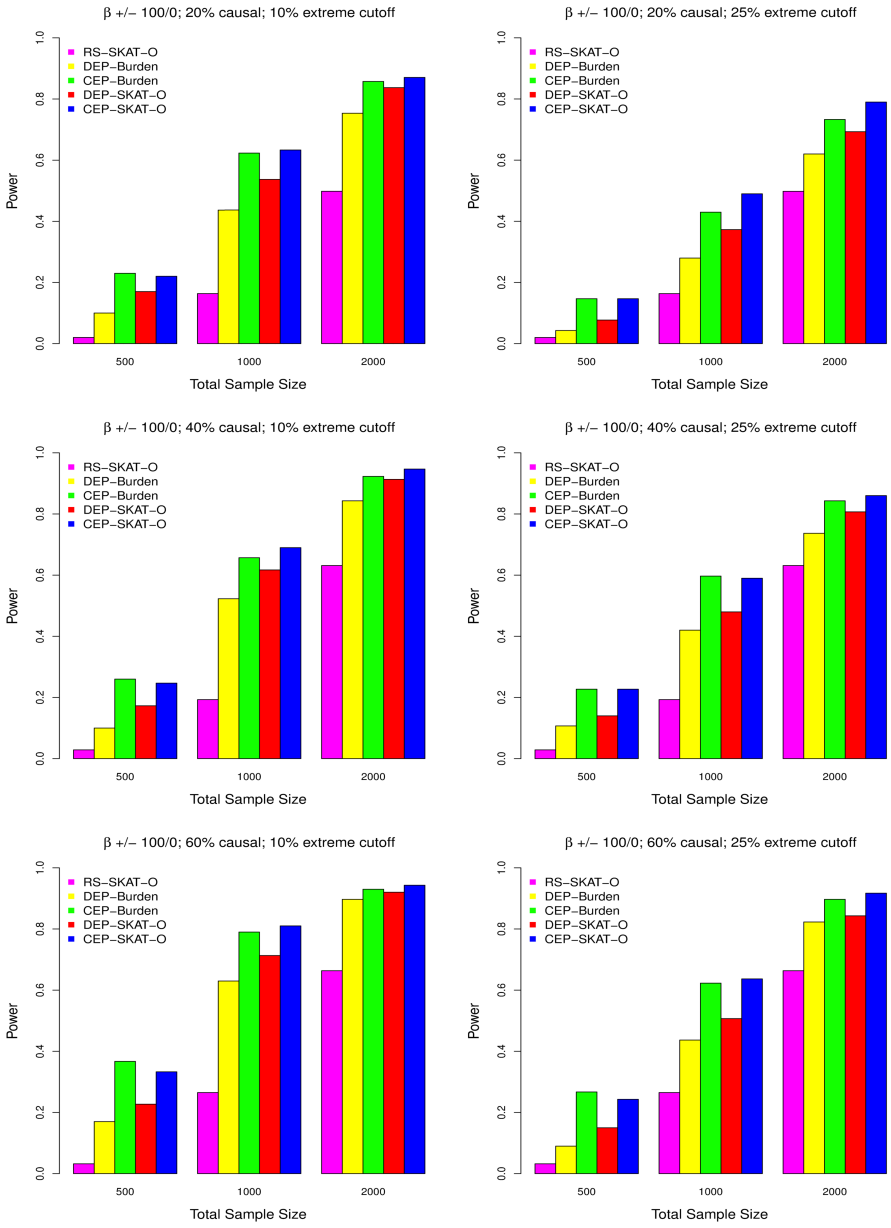

Figure A.3: Simulated power comparisons between four rare variants association tests with all causal variants having a positive effect on phenotype. The five tests are random sample optimal SKAT (RS-SKAT-O), dichotomized extreme phenotype burden test (DEP-Burden), continuous extreme phenotype burden test (CEP-Burden), dichotomized extreme phenotype optimal SKAT (DEP-SKAT-O), and continuous extreme phenotype optimal SKAT (CEP-SKAT-O). The left panel considers the situation where 10% high/low extremes are sampled with the three rows corresponding to 20% (0.6% heritability), 40% (1.2% heritability) and 60% (1.8% heritability) variants in a 3kb region being causal. Three total sample sizes are considered: n=500, 1000, 2000. The right panel considers the situation where 25% high/low extremes are sampled. Exonic regions are simulated with effect sizes for each causal variant equal to $\beta = 1$. Power is estimated by the proportion of tests that detect an association at the $\alpha = 10^{-6}$ level.

Appendix B

HC p-value calculation details and inaccuracy of asymptotic distribution

B.1 Proof of Lemma 1

Proof. Recall $c(t|h) = h[2d\bar{\Phi}(t)\{1 - 2\bar{\Phi}(t)\}]^{1/2} + 2d\bar{\Phi}(t)$. By letting $t \to 0$, we see that $h \geq 0$ and the p-value in this case is 1. Considering a fixed $h > 0$, we evaluate the behavior of $c(t|h)$. This function has several important properties. Firstly, $c(0|h) = d$, $\lim_{t\to\infty} c(t|h) = 0$, and the first derivative is

$$c'(t|h) = d\phi(t)\left(\frac{h(4\bar{\Phi}(t) - 1)}{[2d\bar{\Phi}(t)\{1 - 2\bar{\Phi}(t)\}]^{1/2}} - 2\right).$$

Letting $t_{\max} = \Phi^{-1}[1 - \{1 + d(h^2 + d)^{-1}\}/4]$, we can see from its derivative that $c(t|h)$ is an increasing function on $0 < t < t_{\max}$, achieves its maximum at $t_{\max}$, and then is a decreasing function on $t_{\max} < t < \infty$ approaching 0. This along with $c(0|h) = d$ and $c(t|h)$ being a continuous function gives the result. □

B.2 Proof of Theorem 1

Proof. Firstly, we need only consider $t \geq t_1$ rather than take the intersection over all $t > 0$. This is because for $0 < t < t_1$, $c(t|h) > d$, and as $S(t)$ can never exceed d, $S(t)$ must be less than $c(t|h)$ for every t in this interval with probability 1. Therefore

$$\text{pr}\left[\bigcap_{t>0}\{S(t) < c(t|h)\}\right] = \text{pr}\left[\bigcap_{t\geq t_1}\{S(t) < c(t|h)\}\right].$$

Noting that $S(t)$ an integer-valued non-increasing function of t we have for each k in $1, ..., d$ that

$$\bigcap_{t_k < t \leq t_{k+1}} \{S(t) < c(t|h)\} = \{S(t_k) \leq d - k\}\,.$$

Using this and breaking the intersection into its partition

$$\bigcap_{t \geq t_1} \{S(t) < c(t|h)\} = \bigcap_{k=1}^{d} \bigcap_{t_k < t \leq t_{k+1}} \{S(t) < c(t|h)\}\,,$$

the results follow. □

B.3 Inaccuracy of the asymptotic distribution of the higher criticism in finite d settings

Asymptotically, one needs only take the supremum over a small subset of $t > 0$ (Donoho and Jin, 2004). However, in finite d settings it is necessary to take the supremum over the full $t > 0$ region. This corresponds to $\epsilon = 0$ and $\delta = 1$, and the asymptotic convergence to the gumbel distribution is even slower in this case than what is observed in Fig. B.1. At the $\alpha = 0{\cdot}05$ significance level, the Type I error is $0{\cdot}03$ for $d = 10^4$ if calculated using the asymptotic distribution with $\epsilon = 0{\cdot}01$ and $\delta = 0{\cdot}40$. This inaccurate Type I error rate for even very large d demonstrates how this asymptotic result is not useful for even moderately large d settings. The results are displayed in Figure B.1.

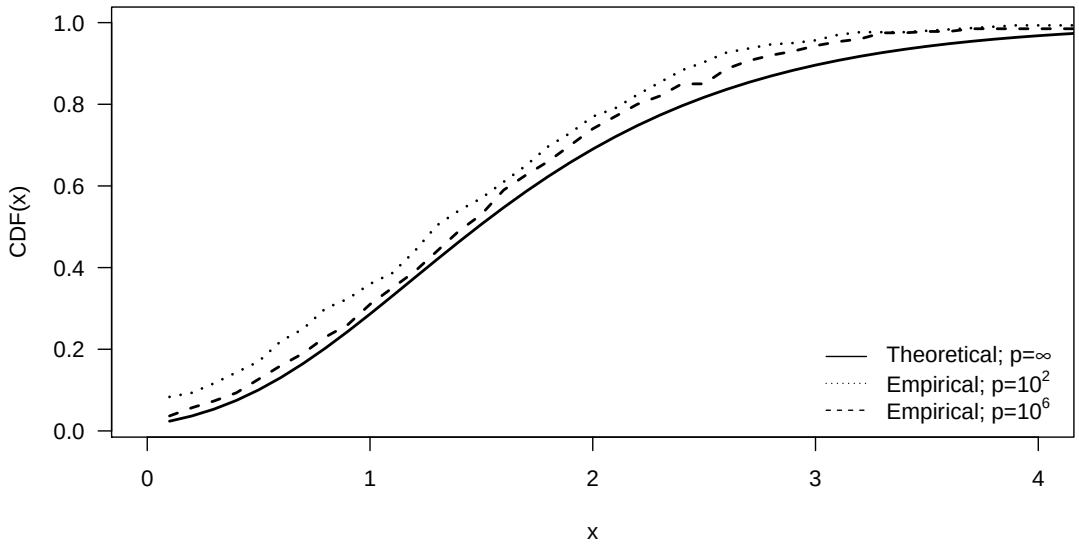

Figure B.1: Comparison of the asymptotic distribution and the empirical distribution of the higher criticism as a function of d. The supremum of the higher criticism test statistic is taken over the range $\Phi^{-1}(1-\delta/2) < t < \Phi^{-1}(1-\epsilon/2)$ where $\epsilon = 0{\cdot}01$ and $\delta = 0{\cdot}40$. Each empirical distribution is constructed using 500 samples of d independent standard normal random variables.

Appendix C

Proofs of GHC detection boundary and p-value calculation

C.1 Proof of Theorem 3

Proof.

$$
\begin{aligned}
&Cov\left(\sum_{k=1}^{p} I_{\{|Z_k|>t_i\}}, \sum_{k=1}^{p} I_{\{|Z_k|>t_j\}}\right) \\
&= E\left((\sum_{k=1}^{p} I_{\{|Z_k|>t_i\}})(\sum_{k=1}^{p} I_{\{|Z_k|>t_j\}})\right) - E\left(\sum_{k=1}^{p} I_{\{|Z_k|>t_i\}}\right) E\left(\sum_{k=1}^{p} I_{\{|Z_k|>t_j\}}\right) \\
&= E\left(\sum_{k=1}^{p} I_{\{|Z_k|>max\{t_i,t_j\}\}} + \sum_{k\neq l} I_{\{|Z_k|>t_i\}} I_{\{|Z_l|>t_j\}}\right) - 4p^2\bar{\Phi}(t_i)\bar{\Phi}(t_j) \\
&= p[2\bar{\Phi}(max\{t_i,t_j\}) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)] + \sum_{k\neq l}[P(|Z_k|>t_i, |Z_l|>t_j) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)]
\end{aligned}
$$

So it is sufficient to show that

$$
\sum_{k\neq l}[P(|Z_k|>t_i, |Z_l|>t_j) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)] = 4p(p-1)\phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{\mathcal{H}_{2n-1}(t_i)\mathcal{H}_{2n-1}(t_j)\overline{r^{2n}}}{(2n)!}
$$

Letting $r_{k,l} = Cov(Z_k, Z_l)$, Schwartzman and Lin (2011) showed that

$$
P(Z_k>t_i, Z_l>t_j) = \bar{\Phi}(t_i)\bar{\Phi}(t_j) + \phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{r_{k,l}^n}{n!}\mathcal{H}_{n-1}(t_i)\mathcal{H}_{n-1}(t_j)
$$

Because Z_k and Z_l are bivariate normal we can rewrite $P(|Z_k| > t_i, |Z_l| > t_j)$ as:

$$P(|Z_k| > t_i, |Z_l| > t_j) = 2(\bar{\Phi}(t_i) - P(Z_k > t_i, Z_l > -t_j) + P(Z_k > t_i, Z_l > t_j))$$

Plugging back in yields:

$$\begin{aligned}
&\sum_{k \neq l} [P(|Z_k| > t_i, |Z_l| > t_j) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)] \\
&= \sum_{k \neq l} 2\phi(t_i)\phi(t_j) \sum_{n=1}^{\infty} \frac{r_{k,l}^n}{n!} \mathcal{H}_{n-1}(t_i)(\mathcal{H}_{n-1}(t_j) - \mathcal{H}_{n-1}(-t_j)) \\
&= 2\phi(t_i)\phi(t_j) \sum_{n=1}^{\infty} \frac{\mathcal{H}_{n-1}(t_i)(\mathcal{H}_{n-1}(t_j) - \mathcal{H}_{n-1}(-t_j))}{n!} \sum_{k \neq l} r_{k,l}^n \\
&= 4p(p-1)\phi(t_i)\phi(t_j) \sum_{n=1}^{\infty} \frac{\mathcal{H}_{2n-1}(t_i)\mathcal{H}_{2n-1}(t_j)\overline{r^{2n}}}{(2n)!}
\end{aligned}$$

C.2 Proof of the GHC P-value calculation

Proof.

$$\begin{aligned}
pr\,(GHC \geq h) &= 1 - pr\left(\bigcap_{t>0} \left\{ S(t) < h\sqrt{\widehat{Var}(S(t))} + 2p\bar{\Phi}(t) \right\}\right) \\
&= 1 - pr\left(\bigcap_{k=1}^{p} \{S(t_k) < p - k + 1\}\right)
\end{aligned}$$

where the t_k are defined in equation (3.5). We are able to write the intersection over all $t > 0$ as an intersection of p events due to the monotone nature of $h\sqrt{\widehat{Var}(S(t))} + 2p\bar{\Phi}(t)$ combined with the fact that $S(t)$ can only take on the values $\{0, 1, ..., p\}$. Applying the

chain rule of conditioning leads to:

$$\begin{aligned} pr\left(GHC \geq h\right) &= 1 - pr\left(\bigcap_{k=1}^{p}\{S(t_k) < p-k+1\}\right) \\ &= 1 - \prod_{k=1}^{p} pr\left(S(t_k) \leq p-k \middle| \bigcap_{l=1}^{k-1}\{S(t_l) \leq p-l\}\right) \\ &= 1 - \prod_{k=1}^{p}\sum_{a=0}^{p-k} q_{k,a} \end{aligned}$$

C.3 Proof of Theorem 4

Proof. Let $\sigma_a(t) = \sqrt{Var(S(t))}$ and $\sigma_s(t) = \sqrt{2p\bar{\Phi}(t)(1-2\bar{\Phi}(t))}$, and then let $HC(t) = \{S(t) - 2p\bar{\Phi}(t)\}/\sigma_s(t)$ and $GHC(t) = \{S(t) - 2p\bar{\Phi}(t)\}/\sigma_a(t)$. Noting that $GHC(t)$ is a mean 0 variance 1 random variable,

$$\begin{aligned} pr_{H_0}\left(GHC > c\right) &\leq \sum_{t\in[s,\sqrt{5\log p}]\cap\mathbb{N}} pr_{H_0}(GHC(t) > c) \\ &\leq \sum_{t\in[s,\sqrt{5\log p}]\cap\mathbb{N}} 1/c^2 \qquad \text{by Chebyshev's Inequality} \\ &= \frac{O(\sqrt{\log p})}{c^2} \end{aligned}$$

Hence for $c = O(\log p)$ we have that $pr_{H_0}(GHC > c) \to 0$. Without loss of generality take $c = \log p$.

Now we study the behavior of GHC under the alternative. By Arias-Castro et al. (2011) we have that if $\max_j |\beta_j| \geq \sqrt{6\log p}$, then

$$HC(\sqrt{5\log p}) \geq p^{3/4} \tag{C.1}$$

with probability greater than $1 - o(1/\sqrt{p})$. For the rest of the alternatives satisfying $A \leq$

$\max_j |\beta_j| \leq \sqrt{6\log p}$, it suffices to show that there exists a $t \in [s, \sqrt{5\log p}] \cap \mathbb{N}$ such that

$$E_{H_1}(GHC(t)) \gg \log p$$

and

$$\frac{E_{H_1}(GHC(t))}{\sqrt{Var_{H_1}(GHC(t))}} \to \infty$$

Noting that $GHC(t) = HC(t)\frac{\sigma_s(t)}{\sigma_a(t)}$, we have that

$$\frac{E_{H_1}(GHC(t))}{\sqrt{Var_{H_1}(GHC(t))}} = \frac{E_{H_1}(HC(t))}{\sqrt{Var_{H_1}(HC(t))}}$$

In Arias-Castro et al. (2011), proof of Theorem 3, they show that for $t = \sqrt{2\min(1, 4\gamma)\log p}$, $\frac{E_{H_1}(HC(t))}{\sqrt{Var_{H_1}(HC(t))}} \to \infty$. Hence, for the same t, $\frac{E_{H_1}(GHC(t))}{\sqrt{Var_{H_1}(GHC(t))}} \to \infty$.

We will show that for that same t, $E_{H_1}(GHC(t)) = \frac{\sigma_s(t)}{\sigma_a(t)} E_{H_1}(HC(t)) \gg \log p$. For the same t, Arias-Castro et al. (2011) show that $E_{H_1}(HC(t)) \gg (\log p)^2\sqrt{\Delta}$. This implies that $E_{H_1}(GHC(t)) \gg \frac{\sigma_s(t)}{\sigma_a(t)}(\log p)^2\sqrt{\Delta}$.

Arias-Castro et al. (2011) showed that $Var_{H_0}(HC(t')) \leq c'(\log p)^2\Delta$ for some constant $c' > 0$. Combine this inequality with the fact that $Var_{H_0}(HC(t')) = \frac{\sigma_a^2(t')}{\sigma_s^2(t')}$, and we have that $\frac{\sigma_s(t)}{\sigma_a(t)} \leq \frac{1}{\sqrt{c'}\log p\sqrt{\Delta}}$. Hence,

$$\begin{aligned} E_{H_1}(GHC(t)) &\gg \frac{1}{\sqrt{c'}\log p\sqrt{\Delta}}(\log p)^2\sqrt{\Delta} \\ &= O(\log p) \end{aligned}$$

Therefore $E_{H_1}(GHC(t)) \gg \log p$ as required. Now we evaluate the case where $t = \sqrt{5\log p}$:

$$\begin{aligned} GHC(\sqrt{5\log p}) &= HC(\sqrt{5\log p})\frac{\sigma_s(\sqrt{5\log p})}{\sigma_a(\sqrt{5\log p})} \\ &\gg p^{3/4}\frac{1}{\log p\sqrt{\Delta}} && \text{by equation (C.1)} \\ &\gg \log p \end{aligned}$$